

# MULTI-MODAL CAPTURING AND SYNTHESIS OF DIGITAL HUMANS

---



A dissertation submitted to  
TECHNISCHE UNIVERSITÄT DARMSTADT  
Fachbereich Informatik

in fulfillment of the requirements for the degree of  
Doktor-Ingenieur (Dr.-Ing.)

presented by  
WOJCIECH ZIELONKA  
M.Sc.

---

Examiner: Prof. Justus Thies, Ph.D.  
Co-examiner: Prof. Matthias Nießner, Ph.D.  
Date of Submission: June 27<sup>th</sup>, 2025  
Date of Defense: August 12<sup>th</sup>, 2025

---

Darmstadt, 2025



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT





*Multi-modal Capturing and Synthesis of Digital Humans*  
*Creation of Digital Humans*

Submitted doctoral thesis by Wojciech Zielonka

Examiner: Prof. Justus Thies, Ph.D.

Co-examiner: Prof. Matthias Nießner, Ph.D.

Date of Submission: June 27<sup>th</sup>, 2025

Date of Defense: August 12<sup>th</sup>, 2025

Darmstadt, Technische Universität Darmstadt

Jahr der Veröffentlichung der Dissertation auf TUPrints: 2025

Bitte zitieren Sie dieses Dokument als:

URN: urn:nbn:de:tuda-tuprints-xxxxx

URL: <https://tuprints.ulb.tu-darmstadt.de/xxxxx>

Dieses Dokument wird bereitgestellt von tuprints,

E-Publishing-Service der TU Darmstadt

<http://tuprints.ulb.tu-darmstadt.de>

[tuprints@ulb.tu-darmstadt.de](mailto:tuprints@ulb.tu-darmstadt.de)

© 2025 Wojciech Zielonka.

This item is protected by copyright and/or related rights. You are free to use this work in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses, you need to obtain permission from the rights-holder(s).

More information about this copyright statement is available at

<http://rightsstatements.org/vocab/InC/1.0/>.



*To my family, who have always believed in me, even when I doubted myself.  
To my friends, whose support and humor carried me through the hardest  
moments. And to my girlfriend, whose love, patience, and understanding  
have meant everything — this journey would not have been the same without  
you.*



## ABSTRACT

---

Multi-modal capture and synthesis of digital humans represent a complex and multifaceted challenge. Achieving realistic digital twins requires addressing a wide range of intricate details. Moreover, humans are highly sensitive to visual artifacts; even minor inconsistencies in facial reconstruction or garment simulation can significantly impair perceived realism. Nevertheless, these virtual doppelgängers play a critical role in applications such as virtual and mixed reality, conversational AI, and virtual conferencing. Over the past two decades, progress in this domain has been exponential, driven by advances in computer vision, computer graphics, machine learning, and deep learning, which have significantly improved the quality and realism of digital human representations. In this thesis, we present a series of projects that have advanced the state of the art in capture and synthesis techniques, enabling the creation of photorealistic avatars that bring us closer to truly realistic digital humans.

Among these, one of the fundamental requirements for virtual and mixed reality applications is the metrically accurate reconstruction of the human head. To enable correct shape capture, we introduce a face reconstruction approach based on a single-shot regression network that recovers metrically accurate 3D head geometry from a single image. Building on this, we present a metrically initialized monocular face tracker that leverages the recovered face shape to improve the tracking results. Furthermore, we propose an extension to existing benchmarks to enable the quantitative evaluation of metric reconstruction accuracy. However, metrical precision is not the only challenge. A key limitation of current avatar methods is their inability to adapt to changes in the appearance of the driving actor. To overcome the constraints of offline-trained, pre-recorded avatars, we introduce a real-time reconstruction and rendering pipeline capable of instantaneously capturing and synthesizing metrically accurate 4D human head avatars. Our method employs neural graphics primitives rigged via deformation gradients computed between canonical and deformed spaces, yielding an efficient representation for both capture and synthesis. Furthermore, the pipeline can support continuous avatar updates, making it particularly well-suited for immersive applications in which avatar quality can be refined in real time through an online data stream.

Realistic full-body avatars require not only the accurate capture of the human head but also the synthesis of fine-grained details, such as pose-driven wrinkles on clothing and self-shadows. To this end, we

introduce an efficient neural representation for fully articulated digital twins that simultaneously encodes facial expressions and garment deformations. Our method embeds Gaussian primitives within per-part tetrahedral cages and leverages deformation transfer by modulating Gaussian kernel parameters to simulate complex phenomena such as stretching and sliding. The resulting model is lightweight, operates in real time, and supports both avatar decomposition and localized conditioning of deformations. However, this method requires long enrollment videos, which are not available in scenarios where the user wants to create an avatar at home. To this end, we propose a few-shot inversion framework that reconstructs a fully controllable 4D head avatar from only a handful of in-the-wild images. Our method leverages a prior network trained exclusively on synthetic data. Through a novel fine-tuning procedure, we demonstrate that, when trained on a sufficiently diverse synthetic dataset, our approach eliminates the need for costly multi-view capture setups while achieving high-quality and robust inversion.

The quality of avatars is often constrained by the capacity of the underlying neural networks. Achieving ultra-realistic capture typically requires large and powerful convolutional architectures. However, such models are often unsuitable for deployment on commodity devices without dedicated hardware accelerators. To overcome this limitation while maintaining high-quality output, we introduce a distillation framework that leverages a pre-trained neural network to extract a lightweight linear eigenbasis representation. This compact model enables high-fidelity face modeling and synthesis on commodity hardware, marking a significant step toward efficient avatar representations that serve as practical alternatives to computationally intensive deep networks. As an application, we develop an image-space cross-reenactment framework that transfers facial expressions from a driving actor to our lightweight avatar in real time.

The work presented in this thesis serves as a foundation for numerous downstream applications. Our contributions advance both the capture and synthesis aspects of digital humans, with a focus on appearance modeling and reconstruction accuracy. This enables applications where the emphasis lies in motion generation, under the assumption that a high-quality human model is provided. In such scenarios, our methods support the correct synthesis of realistic avatars under novel expressions, poses, or viewpoints.

Multi-modale Erfassung und Synthese digitaler Menschen stellen eine komplexe und facettenreiche Herausforderung dar. Die realistische Nachbildung digitaler Zwillinge erfordert die Berücksichtigung zahlreicher feiner Details. Darüber hinaus sind Menschen äußerst empfindlich gegenüber visuellen Artefakten; selbst kleine Unstimmigkeiten bei der Gesichtsrekonstruktion oder der Simulation von Kleidung können die wahrgenommene Realitätsnähe erheblich beeinträchtigen. Dennoch spielen diese virtuellen Doppelgänger eine zentrale Rolle in Anwendungen wie virtueller und gemischter Realität, Konversations-KI und virtuellen Konferenzen. In den letzten zwei Jahrzehnten wurde in diesem Bereich exponentieller Fortschritt erzielt, angetrieben durch Entwicklungen in der Computer Vision, Computergrafik, dem maschinellen Lernen und dem Deep Learning, die die Qualität und Realitätsnähe digitaler menschlicher Darstellungen erheblich verbessert haben. In dieser Arbeit präsentieren wir eine Reihe von Projekten, die den Stand der Technik in der Erfassungs- und Synthesetechnologie vorangetrieben haben und die Erstellung fotorealistischer Avatare ermöglichen, die uns realistischen digitalen Menschen näherbringen.

Eine der grundlegenden Anforderungen für Anwendungen in virtueller und gemischter Realität ist die metrisch genaue Rekonstruktion des menschlichen Kopfes. Um eine korrekte Formaufnahme zu ermöglichen, stellen wir einen Gesichtsrekonstruktionsansatz vor, der auf einem Single-Shot-Regressionsnetzwerk basiert und aus einem einzigen Bild metrisch genaue 3D-Kopfgeometrie rekonstruiert. Darauf aufbauend präsentieren wir einen metrisch initialisierten monokularen Gesichtstracker, der die rekonstruierte Gesichtsform nutzt, um die Tracking-Ergebnisse zu verbessern. Darüber hinaus schlagen wir eine Erweiterung bestehender Benchmarks vor, um eine quantitative Bewertung der metrischen Rekonstruktionsgenauigkeit zu ermöglichen. Allerdings ist die metrische Präzision nicht die einzige Herausforderung. Eine zentrale Einschränkung aktueller Avatar-Methoden ist ihre Unfähigkeit, sich an Veränderungen im Erscheinungsbild des steuernden Akteurs anzupassen. Um die Einschränkungen offline trainierter, vorab aufgezeichneter Avatare zu überwinden, stellen wir eine Echtzeit-Erfassungs- und Renderpipeline vor, die in der Lage ist, metrisch genaue 4D-Kopfavatare sofort zu erfassen und zu synthetisieren. Unsere Methode verwendet neurale Grafikprimitive, die über Deformationsgradienten zwischen kanonischem und deformiertem

---

<sup>1</sup> AUTOMATISCH INS DEUTSCHE ÜBERSETZT.

Raum rigged werden, was eine effiziente Repräsentation für Erfassung und Synthese ermöglicht. Darüber hinaus unterstützt die Pipeline kontinuierliche Avatar-Updates, was sie besonders geeignet für immersive Anwendungen macht, in denen die Avatarqualität durch einen Online-Datenstrom in Echtzeit verbessert werden kann.

Realistische Ganzkörper-Avatare erfordern nicht nur die genaue Erfassung des menschlichen Kopfes, sondern auch die Synthese feiner Details wie posengetriebene Falten in der Kleidung und Selbstschatten. Zu diesem Zweck stellen wir eine effiziente neuronale Repräsentation für vollständig artikulierte digitale Zwillinge vor, die gleichzeitig Gesichtsausdrücke und Kleidungsdeformationen kodiert. Unsere Methode bettet Gaußsche Primitive in teilweise tetraedrische Käfige ein und nutzt Deformationstransfer durch Modulation der Gauß-Kernel-Parameter zur Simulation komplexer Phänomene wie Streckung und Gleitbewegung. Das resultierende Modell ist leichtgewichtig, arbeitet in Echtzeit und unterstützt sowohl die Zerlegung von Avataren als auch die lokal konditionierte Steuerung von Deformationen. Diese Methode erfordert jedoch lange Einschreibevideos, die in Szenarien, in denen der Nutzer zu Hause einen Avatar erstellen möchte, nicht verfügbar sind. Zu diesem Zweck schlagen wir ein Few-Shot-Inversionsframework vor, das einen vollständig steuerbaren 4D-Kopfavatar aus nur wenigen Bildern aus freier Wildbahn rekonstruiert. Unsere Methode nutzt ein Prior-Netzwerk, das ausschließlich auf synthetischen Daten trainiert wurde. Durch ein neuartiges Feinabstimmungsverfahren zeigen wir, dass unser Ansatz, bei ausreichend diverser synthetischer Trainingsmenge, auf kostspielige Multi-View-Erfassungs-Setups verzichten kann und dennoch qualitativ hochwertige und robuste Inversionen erreicht.

Die Qualität von Avataren ist oft durch die Kapazität der zugrunde liegenden neuronalen Netzwerke begrenzt. Die Erreichung ultra-realistischer Erfassung erfordert typischerweise große und leistungsfähige konvolutionale Architekturen. Solche Modelle sind jedoch häufig ungeeignet für den Einsatz auf handelsüblichen Geräten ohne dedizierte Hardwarebeschleuniger. Um diese Einschränkung zu überwinden und dennoch hochwertige Ergebnisse zu erzielen, stellen wir ein Distillationsframework vor, das ein vortrainiertes neuronales Netzwerk nutzt, um eine leichtgewichtige lineare Eigenbasisrepräsentation zu extrahieren. Dieses kompakte Modell ermöglicht hochqualitative Gesichtsmodellierung und -synthese auf handelsüblicher Hardware und stellt damit einen bedeutenden Schritt hin zu effizienten Avatar-Repräsentationen dar, die als praktikable Alternativen zu rechenintensiven Deep-Learning-Netzwerken dienen. Als Anwendung entwickeln wir ein image-space Cross-Reenactment-Framework, das Gesichtsausdrücke von einem steuernden Akteur in Echtzeit auf unseren leichtgewichtigen Avatar überträgt.



Die in dieser Arbeit präsentierten Beiträge bilden eine Grundlage für zahlreiche weiterführende Anwendungen. Unsere Arbeiten treiben sowohl die Erfassung als auch die Synthese digitaler Menschen voran, mit einem Fokus auf Erscheinungsmodellierung und Rekonstruktionsgenauigkeit. Dies ermöglicht Anwendungen, bei denen die Bewegungsgenerierung im Vordergrund steht, unter der Annahme, dass ein qualitativ hochwertiges menschliches Modell gegeben ist. In solchen Szenarien ermöglichen unsere Methoden die korrekte Synthese realistischer Avatare unter neuen Ausdrücken, Posen oder Blickwinkeln.



## PUBLICATIONS

---

**Wojciech Zielonka**, Timo Bolkart, Justus Thies. "Towards Metrical Reconstruction of Human Faces" In: European Conference on Computer Vision (ECCV) 2022.

Website: <https://zielon.github.io/mica/>

**Wojciech Zielonka**, Timo Bolkart, Justus Thies. "Instant Volumetric Head Avatars" In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2023.

Website: <https://zielon.github.io/insta/>

**Wojciech Zielonka**, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, Javier Romero. "Drivable 3D Gaussian Avatars" In: International Conference on 3D Vision (3DV) 2025.

Website: <https://zielon.github.io/d3ga/>

**Wojciech Zielonka**, Timo Bolkart, Thabo Beeler, Justus Thies. "Gaussian Eigen Models for Human Heads" In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025.

Website: <https://zielon.github.io/gem/>

**Wojciech Zielonka**, Stephan J. Garbin, Alexandros Lattas, George Kopanas, Paulo Gotardo, Thabo Beeler, Justus Thies, Timo Bolkart. "Synthetic Prior for Few-Shot Drivable Head Avatar Inversion" In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025.

Website: <https://zielon.github.io/synshot/>

Berna Kabadayi, **Wojciech Zielonka**, Bharat Lal Bhatnagar, Gerard Pons-Moll, Justus Thies. "Controllable Personalized GAN-based Human Head Avatar" In: International Conference on 3D Vision (3DV) 2024.

Website: <https://zielon.github.io/ganavatar/>



## ACKNOWLEDGMENTS

---

First and foremost, I would like to thank my supervisor and mentor, Justus Thies, who has guided me since my master’s studies. He introduced me to neural rendering, face modeling, and tracking, and ignited my passion for research and discovery in computer vision and graphics. I am also deeply grateful to Timo Bolkart, with whom I collaborated throughout my PhD; his insightful advice, attention to detail, and creative ideas have substantially improved many of my papers. I would like to thank Javier Romero for hosting me during my industry internship. During that time, I learned a great deal, which enhanced the quality of my research and helped me grow as a scientist. I would also like to thank my office mates, whose camaraderie and support made long working hours far more bearable and enjoyable, especially during the most demanding deadlines. Finally, I extend my deepest gratitude to the Max Planck Institute for Intelligent Systems for providing a unique environment in which researchers can tackle challenging problems without concern for computational resources, a support that is particularly invaluable in today’s era of artificial intelligence.



## CONTENTS

---

1	INTRODUCTION	1
2	BACKGROUND	7
2.1	3D Morphable Models . . . . .	7
2.2	Human Head Avatars . . . . .	13
2.3	Full-body Avatars . . . . .	15
2.4	Generative Modeling of Humans . . . . .	17
3	PUBLICATIONS AND CONTRIBUTIONS	21
3.1	Metrical Accurate Human Shape Regression . . . . .	21
3.2	Instant Head Avatar Creation . . . . .	25
3.3	Full-body 3D Gaussian Avatars . . . . .	28
3.4	Distillation of Avatars into a Linear Model . . . . .	31
3.5	Building a Synthetic Prior for Few-Shot Inversion . . . . .	34
4	DISCUSSION	37
4.1	Summary of Contributions . . . . .	37
4.2	Potential Limitations . . . . .	39
4.3	Future Work . . . . .	40
4.4	Conclusions . . . . .	41
A	APPENDIX	43
A.1	Towards Metrical Reconstruction of Human Faces . . . . .	43
A.2	Instant Volumetric Head Avatars . . . . .	66
A.3	Drivable 3D Gaussian Avatars . . . . .	78
A.4	Gaussian Eigen Models for Human Heads . . . . .	91
A.5	Synthetic Prior for Few-Shot Drivable Head Avatar In- version . . . . .	103
A.6	Broader Impact: Ethical Concerns . . . . .	116
	BIBLIOGRAPHY	117

## LIST OF FIGURES

---

Figure 1.1	This thesis introduces several publications (MICA [233], INSTA [234], D3GA [231], GEM [232], SynShot [235]) that tackle essential stages in the creation of digital humans: capturing and synthesis. . . . .	1
Figure 1.2	<b>Geometric Capturing</b> involves obtaining a 3D representation from single or multi-view camera setups. Given a calibrated multi-view setup, ToFu [96] produces topologically consistent meshes using a volumetric representation. In in-the-wild scenarios, where only a single image is available, MICA [233] regresses metrically plausible human head reconstructions. . . . .	2
Figure 1.3	Modeling transforms captured data into a parametric representation that can be manipulated, sampled, and animated. FLAME [95] and SMPL [108] model only coarse shape geometry. GEM [232] extends this representation to the personalized case and jointly models both geometry and appearance. . . . .	3
Figure 1.4	<b>Photorealistic Synthesis</b> focuses on generating entirely new faces or bodies, producing realistic appearance conditioned on expressions, either through powerful generative prior models (e.g., Cap4D [175], SynShot [235]) or via personalized models trained on a single actor (e.g., MVP [106], INSTA [234]). . . . .	4
Figure 1.5	Embodied Conversational Agents are digital humans equipped with multimodal reactive capabilities, enabling them to act and move in an authentic and human-like manner. For example, <i>Audio to Photoreal Embodiment</i> [122] presents a framework for generating full-body, photorealistic avatars that gesture in accordance with the conversational dynamics of dyadic interactions. . . . .	5



Figure 2.1	Parametrization of FLAME [95]. Left: Variation of the first three shape principal components across the range of $-3$ to $+3$ standard deviations. Middle: Pose-induced deformations resulting from the rotation of four out of six neck and jaw joints. Right: Variation of the first three expression principal components across the range of $-3$ to $+3$ standard deviations. . .	8
Figure 2.2	3D face reconstruction and tracking constitute a fundamental component in the development of digital avatars. Often, before neural representations can be effectively employed, high-quality tracked datasets are required, providing meshes, appearance information, and other modalities that can be leveraged in learning-based approaches. The figure is courtesy of Zollhöfer <i>et al.</i> [236]. . . . .	10
Figure 2.3	Overview of a face tracking pipeline proposed by Thies <i>et al.</i> [182], which is based on a 3D Morphable Model [13] and enables real-time reconstruction and tracking of human faces. .	11
Figure 2.4	Neural rendering encompasses a broad class of techniques for tasks such as novel-view synthesis of static and dynamic scenes, generative object modeling, and photometric scene relighting. Its popularity in digital avatar creation stems from its ability to learn photorealistic human appearance models directly from data. The illustration is courtesy of Tewari <i>et al.</i> [177, 178]. . . . .	13
Figure 2.5	Modeling full-body avatars is a highly challenging task, as it involves capturing complex aspects such as body motion, garments, hair, and facial expressions. The figure is courtesy of [65, 98, 132, 172]. . . . .	15

## LIST OF TABLES

---

Table 3.1	Average photometric errors over 19 videos spanning our dataset and the public NHA, IMAvatar, and NeRFace benchmarks. “Time” denotes the average rendering time per frame. Our method matches NeRFace on pixel-wise metrics, achieves low perceptual error, and is substantially faster to train and evaluate. . . .	26
Table 3.2	On our dataset, D3GA achieves the highest PSNR and SSIM compared to BodyDecoder [4] and MVP [106]. Among MLP-based avatars, D3GA also leads in image quality—only Animatable Gaussians (AG), with its larger CNN backbone, produces slightly sharper results. .	29
Table 3.3	Quantitative evaluation on novel expressions and views across 16 cameras. GEM, driven by analysis-by-synthesis fitting, outperforms all baselines in PSNR, LPIPS, SSIM, and L1 error.	32

## INTRODUCTION

---

The creation of digital humans is a complex, interdisciplinary domain grounded in computer vision, computer graphics, machine learning, and human-computer interaction. These technologies converge to produce realistic, controllable, and responsive virtual representations of people, with applications ranging from immersive communication and virtual production to gaming, healthcare, and digital assistants. Beyond technical achievements, the development of artificial humans also evokes profound philosophical and ethical reflections. Existentialism emphasizes the individual creation of meaning; creating avatars may reflect our existential drive to define or preserve the self [156]. From a theological perspective, the concept that humans are created “in the image of God” (*imago Dei*), as stated in Genesis 1:26–27 [180], implies that by creating digital replicas in our own image, we assume a quasi-divine role. In the scientific realm, Richard Feynman famously stated, “What I cannot create, I do not understand” [43], emphasizing the epistemic value of creation as a form of understanding. Together, these views invite deeper inquiry into the purpose, identity, and ethical responsibilities involved in the creation of digital humans. To see where these philosophical and ethical dilemmas meet lines of code, we must first unpack the technical pipeline that underlies every digital avatar.

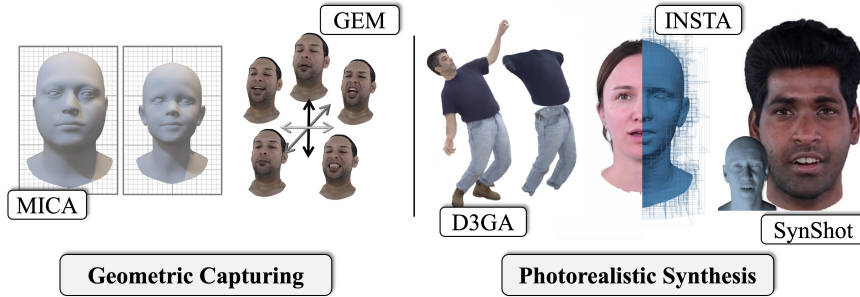


Figure 1.1: This thesis introduces several publications (MICA [233], INSTA [234], D3GA [231], GEM [232], SynShot [235]) that tackle essential stages in the creation of digital humans: capturing and synthesis.

This pipeline can be broadly divided into two stages: capturing and synthesis (Figure 1.1). Each of these stages presents unique challenges and involves distinct technical problems. The capturing stage encompasses appearance estimation, tracking, and reconstruction of human shape and motion. The goal is to obtain either a new model or coefficients for an existing one that can capture the distribution of the human body, facial geometry, and appearance. For example, statistical

models such as FLAME [95], SMPL [108], and machine learning-based generative models [77] provide widely used priors for digital humans. The synthesis stage involves generating photorealistic representations conditioned on motion and appearance. Intricate details of the human face and characteristic garment deformations are essential to ultimately produce a lifelike digital avatar. Finally, all of these stages converge in downstream applications such as photorealistic embodied conversational agents [122] or intelligent virtual agents [109], completing the pipeline for creating artificial humans that not only resemble but also behave like real people.

**Geometric Capturing** humans is a challenging process that often requires multi-view calibrated camera setups to achieve high-quality results [8, 9]. Many 4D avatar methods [54, 175, 202, 208, 232, 234] rely on geometric data for both training and testing, making accurate geometry acquisition essential for realistic appearance synthesis. Figure 1.2 (left) illustrates a volumetric tracking and reconstruction method for human faces, which produces high-quality registered meshes using a regression-based approach that can subsequently be converted into FLAME [95] coefficients. For the problem of in-the-wild reconstruction from monocular images, where camera parameters are unknown (Figure 1.2, right), MICA [233] provides a robust alternative. This method leverages information from a face recognition network [31] by mapping its feature space into the FLAME coefficient space, significantly outperforming other approaches for human shape reconstruction. Since capturing real humans in multi-view setups is time-consuming, alternative strategies that leverage large-scale 2D datasets and self-supervision have gained popularity [33, 39, 55]. Another appealing solution is synthetic data [157, 198, 235], which offers a fully controllable environment; however, many such approaches still struggle with the sim-to-real domain gap.

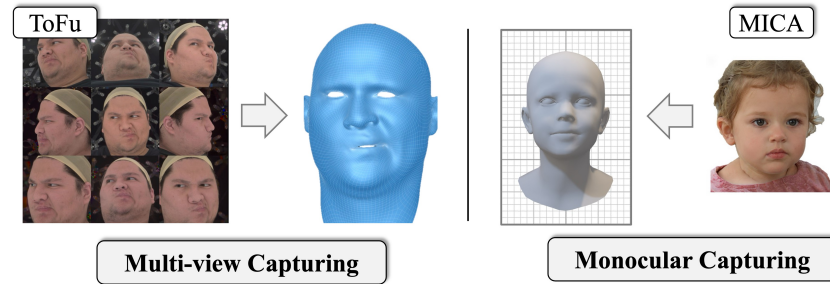


Figure 1.2: **Geometric Capturing** involves obtaining a 3D representation from single or multi-view camera setups. Given a calibrated multi-view setup, ToFu [96] produces topologically consistent meshes using a volumetric representation. In in-the-wild scenarios, where only a single image is available, MICA [233] regresses metrically plausible human head reconstructions.

A fundamental aspect of geometric capturing is the representation and formulation of the model that describes target characteristics such as shape or expressions. This problem is referred to as modeling, and it may involve classical machine learning techniques such as principal component analysis (PCA), as well as more advanced methods including generative neural networks or hybrid approaches that combine the strengths of both paradigms. Figure 1.3 illustrates how models such as SMPL [108] or FLAME [95] provide only coarse approximations of shape. GEM [232], on the other hand, models both geometry and appearance using Gaussian primitives represented as an eigenbasis. Subsequent methods that build upon SMPL or FLAME, e.g., [113, 150, 152, 154, 205, 206, 227, 234], extend their capabilities to capture clothing and hair, thereby delivering significantly greater realism and enhanced modeling flexibility.

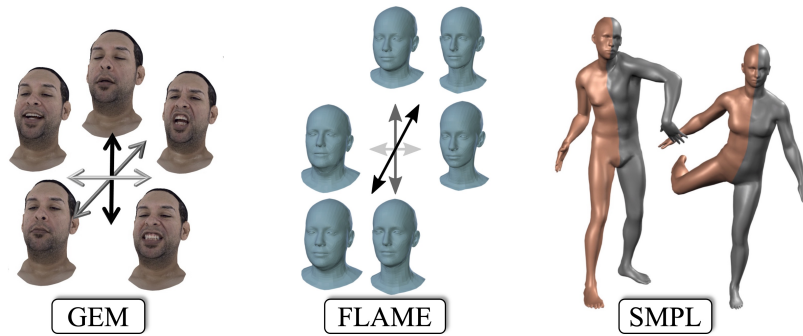


Figure 1.3: Modeling transforms captured data into a parametric representation that can be manipulated, sampled, and animated. FLAME [95] and SMPL [108] model only coarse shape geometry. GEM [232] extends this representation to the personalized case and jointly models both geometry and appearance.

**Photorealistic Synthesis** involves generating realistic virtual humans that exhibit lifelike appearance conditioned on motion or expressions. For example, few-shot inversion methods [86, 157, 175, 209, 225, 235] enable high-quality, controllable avatar synthesis from only a handful of input images. A user can upload a few photographs to the system and, within moments, obtain a 4D avatar suitable for virtual and mixed-reality environments. Over the past decade, this field has progressed from simple, low-frequency PCA-based appearance models [11–13, 182] to highly realistic avatars [54, 106, 151, 232]. Figure 1.4 illustrates several modern methods that either leverage multi-view prior models or adopt personalized architectures to generate high-fidelity avatars. INSTA [234] rigs neural graphics primitives [119] by applying deformation gradients to the primitives. MVP [106], on the other hand, utilizes a personalized VAE to regress volumetric primitives attached to the surface of a mesh, which are later ray-traced and integrated into a final image via volumetric rendering. SynShot [235] builds a generative model using synthetic datasets only

and, through pivotal fine-tuning, adapts the network to individual subjects to achieve high-quality avatars. Cap4D [175] employs multi-view diffusion models to construct a 4D avatar from only four input images. Those approaches demonstrate how a powerful prior can be distilled into a personalized avatar and serve as the backbone for downstream methods [54, 142, 163, 202, 234].

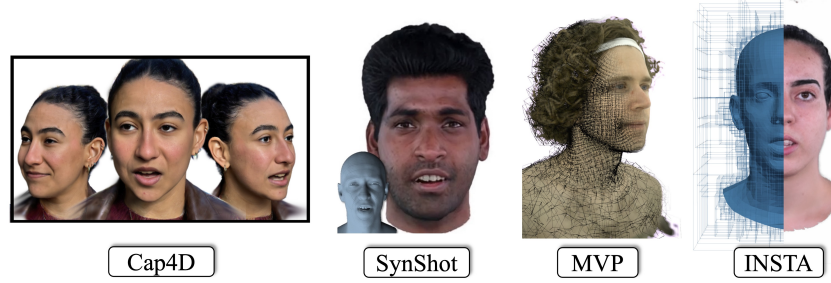


Figure 1.4: **Photorealistic Synthesis** focuses on generating entirely new faces or bodies, producing realistic appearance conditioned on expressions, either through powerful generative prior models (e.g., Cap4D [175], SynShot [235]) or via personalized models trained on a single actor (e.g., MVP [106], INSTA [234]).

Once controllable digital human models are obtained, numerous compelling downstream applications become possible. One particularly interesting and increasingly popular domain is that of Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics, which focuses on how these agents interact with humans through dialogue, multimodal communication, and adaptive behavior. This domain also explores applications in education, healthcare, aging, autism support, and storytelling [109]. Figure 1.5 presents one of the first works in the photorealism direction: Ng et al. [122] developed a framework for generating full-body, photorealistic avatars that gesture according to the conversational dynamics of dyadic interactions. Future directions include exploring motion control [134, 168–170] via reinforcement learning [127] and decision transformers [25, 190] for human motion synthesis. These reasoning capabilities would enable the development of autonomous virtual agents capable of exploring, learning, and interacting—both with one another and with humans in virtual meta-worlds. Potential applications include virtual teaching assistants, telepresence systems, interactive video games, immersive entertainment experiences and many more.

In summary, this thesis investigates the geometric capture and photorealistic synthesis stages of the digital human creation pipeline. The first project, MICA [233] (Section 3.1), addresses face reconstruction within the **geometric capturing** stage: from a single in-the-wild portrait, our pipeline reconstructs a metrically accurate 3D face model suitable for VR/MR applications, where preserving real-world scale is crucial. In Section 3.4, we present our GEM model [232], also part of



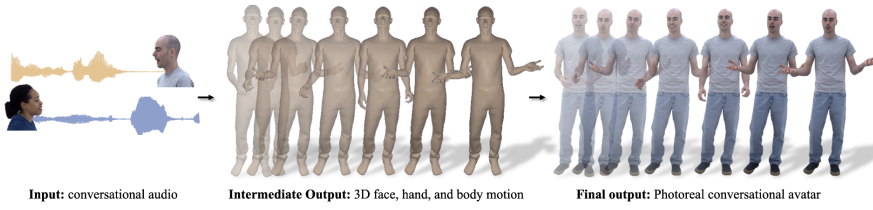


Figure 1.5: Embodied Conversational Agents are digital humans equipped with multimodal reactive capabilities, enabling them to act and move in an authentic and human-like manner. For example, *Audio to Photoreal Embodiment* [122] presents a framework for generating full-body, photorealistic avatars that gesture in accordance with the conversational dynamics of dyadic interactions.

the **geometric capturing** stage. Given a high-fidelity head avatar regressor, we derive a lightweight linear eigenbasis representation, akin to FLAME [95], by applying PCA to the frames regressed from individual training sequences, effectively distilling the neural network into a single-layer linear model. The remaining three projects (Sections 3.3, 3.2, and 3.5) fall under the **photorealistic synthesis** stage. While each addresses a different challenge, their unified goal is the photorealistic and controllable creation of face or full-body avatars given a controlling signal. INSTA [234] (Section 3.2) enables instant creation and real-time rendering of personalized head avatars. D3GA [231] (Section 3.3) generates full-body avatars driven by joint-angle vectors from an underlying parametric body model. Finally, SynShot [235] (Section 3.5) reconstructs photorealistic 4D avatars from as few as three in-the-wild images, despite its prior being trained exclusively on synthetic data.

The motivation for this work is to develop a holistic system capable of generating virtual human avatars that are appearance-indistinguishable from real humans. While we address every stage of the avatar creation pipeline, human appearance remains an extraordinarily complex domain. Challenges such as hair modeling, tongue articulation, garment simulation, and motion synthesis are still only partially solved. Although recent years have brought significant progress in these areas, the creation of fully realistic avatars remains an open problem. The contributions presented in this thesis represent a step toward that goal. In the following chapters, we provide a thorough overview of the necessary background before detailing each project included in this work.





## BACKGROUND

---

### CONTENTS

2.1	3D Morphable Models . . . . .	7
2.2	Human Head Avatars . . . . .	13
2.3	Full-body Avatars . . . . .	15
2.4	Generative Modeling of Humans . . . . .	17

---

Creation of digital avatars is a field that combines a wide range of disciplines, from traditional computer graphics [19, 93, 182] and computer vision to neural rendering [44, 58, 106, 222, 224, 227, 234], generative modeling [1, 73, 126, 235], and, more recently, diffusion models [75, 83, 84]. It is a vast and interdisciplinary area that demands expert knowledge across multiple domains to create digital twins that are indistinguishable from reality [54, 151]. Despite significant progress in recent years, the field still faces many challenging problems, such as realistic hair modeling [164, 216], capturing complex social interactions [109, 122], and achieving robust generalization [28, 209, 225, 235]. This section covers the fundamental computer vision concepts used in the work on digital humans. We begin with 3D Morphable Models (3DMM) [13, 95], followed by face reconstruction [33, 39, 198, 233] and tracking [20, 151, 181, 182, 233, 236]. Next, we describe appearance representation using 3D Gaussian Splatting (3DGS) [80] and Neural Radiance Fields (NeRF) [117] for human heads and full-body avatars. We conclude this chapter with a comprehensive overview of generative modeling techniques as applied to digital humans, highlighting recent advances in diffusion models, rectified flow, and avatar generation [23, 102, 148]. The following sections are adapted from portions of the author’s cumulative work, including the author’s accepted conference papers of [231–235].

### 2.1 3D MORPHABLE MODELS

Eigenfaces [136] represent the first 2D Morphable Model for human faces, computed on grayscale images using Principal Component Analysis (PCA), and were applied to tasks such as face recognition. Later,

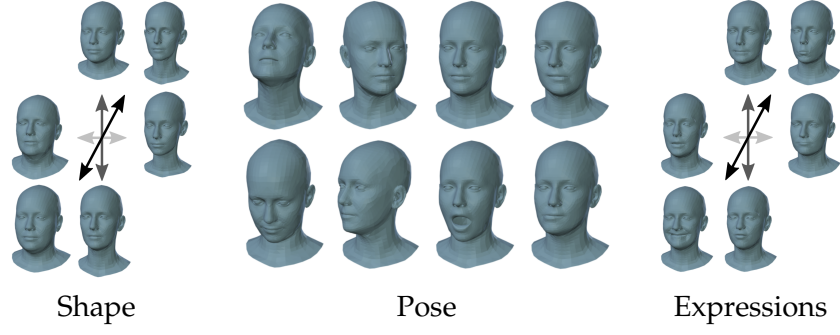


Figure 2.1: Parametrization of FLAME [95]. Left: Variation of the first three shape principal components across the range of  $-3$  to  $+3$  standard deviations. Middle: Pose-induced deformations resulting from the rotation of four out of six neck and jaw joints. Right: Variation of the first three expression principal components across the range of  $-3$  to  $+3$  standard deviations.

Blanz and Vetter [13] introduced the 3D Morphable Model (3DMM) in 1999. Their model was constructed by applying PCA to approximately 200 laser-scanned faces, all aligned to a common template. The resulting eigenvectors capture the principal modes of variation in both geometry and albedo. New face shapes and textures can be synthesized by selecting coefficient vectors  $\delta$ ,  $\sigma$ , and  $\gamma$ , computing linear combinations of the identity, color, and expression bases, and adding them to the respective means:

$$\begin{aligned} \mathbf{S} &= \bar{\mathbf{S}} + \delta \mathbf{B}_{\text{id}} + \gamma \mathbf{B}_{\text{expr}} \\ \mathbf{C} &= \bar{\mathbf{C}} + \sigma \mathbf{D}_{\text{id}} \end{aligned} \quad (2.1)$$

Here,  $\bar{\mathbf{S}}$  is the mean shape,  $\mathbf{B}_{\text{id}}$  and  $\mathbf{B}_{\text{expr}}$  are the identity and expression basis matrices, and  $\delta$  and  $\gamma$  are their corresponding coefficient vectors.  $\mathbf{C}$  represents the low-frequency statistical color texture, with  $\bar{\mathbf{C}}$  as its mean and  $\mathbf{D}_{\text{id}}$  as the color basis. The FLAME model [95] (Figure 2.1) extends the Basel Face Model (BFM) [13] by incorporating linear blend skinning (LBS) for realistic head rotation, along with pose-dependent corrective offsets to capture neck articulation and eyeball rotations.

Since then, many extensions and modifications have been introduced to traditional PCA-based models. For instance, localized models [7, 30, 176] were proposed, using manually selected regions for segmentation. Later, Neumann *et al.* [121] introduced the use of sparse PCA combined with a group sparsity constraint to identify localized deformation components.

The expression space of the Basel Face Model (BFM) is constructed as an additive offset, typically expressed as  $\gamma \mathbf{B}_{\text{expr}}$ , where  $\gamma$  denotes the expression coefficients and  $\mathbf{B}_{\text{expr}}$  represents the linear expression basis. This formulation enables expression transfer between subjects, but it also introduces a significant limitation: the linear nature of the

PCA-derived basis restricts the model’s capacity to capture complex deformations, particularly around the lips and jaw. As a result, the expressiveness of such models is often inadequate for highly dynamic facial motions. To overcome this limitation, several approaches have incorporated nonlinear modeling techniques. For instance, FLAME [95] combines linear expression blendshapes with articulated jaw motion to form a nonlinear expression space that more faithfully captures anatomical constraints and motion. Ichim *et al.* [68] propose a biomechanically inspired muscle activation model, where expressions are driven by physical simulation of facial musculature. Koppen *et al.* [87] adopt a probabilistic approach by modeling both facial geometry and appearance using a Gaussian mixture model. While these methods significantly improve the expressiveness and realism of the resulting face models, they often do so at the cost of increased model complexity and computational overhead. Another line of work employs implicit representations to model expressions. Neural parametric models [125], trained on monocular depth sequences of the human body, learn to embed motion into a latent space that is used to represent both shape and motion. Giebenhain *et al.* [52] apply the same concept to human heads, significantly outperforming traditional linear mesh-based models. For a comprehensive overview of 3D face modeling techniques, we refer the reader to the survey by Egger *et al.* [36].

### 2.1.1 Face Reconstruction

Reconstructing human faces and heads from monocular RGB, RGB-D, or multi-view data is a well-established and actively studied area situated at the intersection of computer vision and computer graphics (Figure 2.4). An extensive overview of optimization-based reconstruction techniques. Particularly, those grounded in the analysis-by-synthesis paradigm is provided by Zollhöfer *et al.* [236], who comprehensively survey the state of the art in this domain. Monocular reconstruction approaches often depend on strong priors on facial shape and appearance to resolve the inherent ambiguity of recovering 3D geometry from a single 2D image under unknown camera transformations [11, 12, 47, 48, 81, 182–187, 196, 197]. These priors are typically encoded in the form of parametric face models or statistical representations, which are optimized to reproduce the observed image through differentiable rendering or direct synthesis. In addition to optimization-based methods, a rich body of work has emerged that leverages learning-based regression techniques to directly infer facial geometry and appearance from input images. A thorough survey of these regression-based approaches—spanning both supervised and self-supervised paradigms—is presented by Morales *et al.* [118], highlighting key trends and challenges in the field.



Figure 2.2: 3D face reconstruction and tracking constitute a fundamental component in the development of digital avatars. Often, before neural representations can be effectively employed, high-quality tracked datasets are required, providing meshes, appearance information, and other modalities that can be leveraged in learning-based approaches. The figure is courtesy of Zollhöfer *et al.* [236].

The advent of the 3D morphable model (3DMM) for human faces by Blanz *et al.* [13] marked a significant milestone in face reconstruction, introducing an optimization-based methodology grounded in the principle of analysis-by-synthesis. In their work, color reproduction was optimized using a sparse sampling strategy, which, while effective, limited the density of the recovered facial details. Subsequently, Thies *et al.* [182, 186] extended this framework by incorporating a dense photometric term that leverages the full facial region, made possible through differentiable rendering techniques applied to the 3DMM. This advancement enabled a range of high-fidelity applications, including the reconstruction of realistic avatars from a single image, even capturing complex structures such as hair [66], as well as the reconstruction of detailed facial reflectance and geometry from unconstrained imagery [211]. Beyond static heads, these methods have been generalized to support full upper-body reconstruction and animation [187], and have been used to model avatars with temporally varying textures [120]. More recently, classical optimization pipelines have been augmented with learnable modules, such as per-vertex surface offsets or view-dependent neural radiance fields, thereby increasing their expressivity and adaptability [58]. Moreover, optimization-based face models often serve as a crucial foundation for modern neural rendering pipelines, including applications such as deep video portraits [81], deferred neural rendering [185], and neural voice puppetry [184], where accurate geometry and appearance priors are essential for photorealistic synthesis and animation.

### 2.1.2 Optimization-based Human Face Tracking

In recent years, numerous novel methods have been proposed for the photorealistic creation of human avatars [54, 142, 208, 223, 234]. However, all of these approaches require some form of control signal.

While some methods employ learned expression encoders [114, 151], the majority rely on 3D Morphable Model (3DMM) expression vectors. To provide these, a robust and real-time 3DMM-based face tracker is essential. Such trackers estimate expression parameters either using sparse facial landmarks [13, 198], dense photometric loss [182, 186], or regression-based methods [33, 39], enabling accurate and efficient expression tracking.

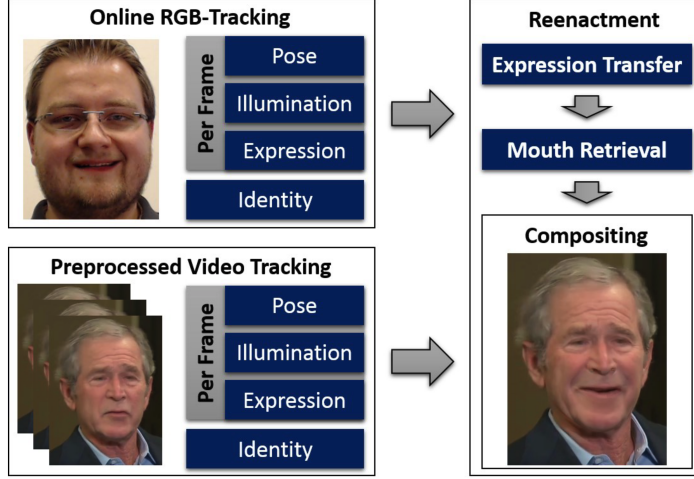


Figure 2.3: Overview of a face tracking pipeline proposed by Thies *et al.* [182], which is based on a 3D Morphable Model [13] and enables real-time reconstruction and tracking of human faces.

Face2Face [182] is a groundbreaking real-time facial reenactment system that performs dense photometric optimization using a custom GPU implementation of a second-order Gauss-Newton optimizer (Figure 2.3). The employed energy formulation follows the principle of *differentiable shading*, which involves sampling both the rasterized and the ground-truth image and applying an  $\mathcal{L}_2$  loss, as described in Equation 2.3. This contrasts with soft rasterizers [88, 90, 103], which implement continuous and *differentiable visibility* testing, as expressed in Equation 2.2. In contrast, traditional pipelines such as OpenGL or Vulkan perform non-differentiable binary visibility tests, which are commonly used in sampling-based approaches.

$$E_{\text{image}}(P) = \sum_{x,y} |I(x,y) - R(x,y,P)| \quad (2.2)$$

$$E_{\text{sample}}(P) = \sum_S |I(x_S, y_S) - C_S(P)| \quad (2.3)$$

This distinction changes the gradient computation, as the sampling-based formulation requires image-space derivatives:

$$\frac{\partial I}{\partial x_S} = \nabla_{x_S} I(x_S, y_S), \quad \frac{\partial I}{\partial y_S} = \nabla_{y_S} I(x_S, y_S) \quad (2.4)$$

In contrast, the image-based formulation requires computing the derivatives of the rendering function  $R(\cdot)$  with respect to the rendering parameters  $P$ :

$$\frac{\partial R}{\partial P} = \nabla_P R(x, y, P) \quad (2.5)$$

In summary, differentiable shading offers a more versatile approach, as it enables not only the optimization of geometry and camera parameters, but also of materials and lighting components that together define the full image formation process. Consequently, methods based on Face2Face [182] are well-suited for estimating 3DMM expression coefficients, which can be subsequently used to control digital avatars at test time in a robust and real-time manner.

### 2.1.3 Regression-based Reconstruction of Human Faces.

Learning-based methods for facial reconstruction can broadly be divided into two categories: supervised and self-supervised approaches. Supervised techniques commonly rely on synthetic renderings of human faces, enabling the training of regressors that predict the parameters of a 3D morphable model (3DMM) from image data [35, 145, 146, 179]. These synthetic datasets provide access to ground truth 3DMM parameters, which allow for direct supervision during training. In contrast, Genova *et al.* [51] propose a hybrid strategy that incorporates both synthetic and real-world data. While synthetic images facilitate supervised learning, real images are used to impose multi-view consistency losses, enforcing that the predicted 3DMM parameters are consistent across different images of the same identity. Their method uses FaceNet [159] embeddings to guide identity preservation in the learned latent space. The DECA model [39] builds upon RingNet [155] by predicting expression-dependent surface offsets in UV space, enhancing the reconstruction of facial deformations. It is trained using dense photometric self-supervised losses applied to in-the-wild images, improving both geometry and appearance fidelity. A similar decomposition strategy, separating coarse geometry (from the 3DMM) and detailed surface variations (via bump or displacement maps), was earlier introduced by Tran *et al.* [188]. Chen *et al.* [24] further extend this idea by combining supervised learning from synthetic renderings with self-supervised objectives to infer both texture and displacement maps from single-view images. Deng *et al.* [33] propose a framework



trained with hybrid-level losses, incorporating multi-image consistency constraints, photometric reconstruction losses with skin-specific attention masks, and perceptual losses guided by FaceNet [159] embeddings. In parallel, Generative Adversarial Network (GAN)-based methods have emerged to predict high-frequency details. These include approaches for estimating detailed UV-space color maps [49, 50] and physical skin properties such as albedo, normals, and specular reflectance [91, 92, 153, 211]. While these works focus on achieving photorealism and expressive detail, the present work is instead focused on reconstructing metrically accurate 3D face representations, prioritizing correctness of spatial scale and geometry over fine-scale texture fidelity. A key challenge in self-supervised monocular methods is the inherent depth-scale ambiguity: the absolute face size, its distance to the camera, and the perspective projection can all vary in ways that produce similar 2D observations. As a result, such models often predict faces at an incorrect scale. This limitation persists even though 3DMMs are inherently defined in a metrically calibrated space.

## 2.2 HUMAN HEAD AVATARS

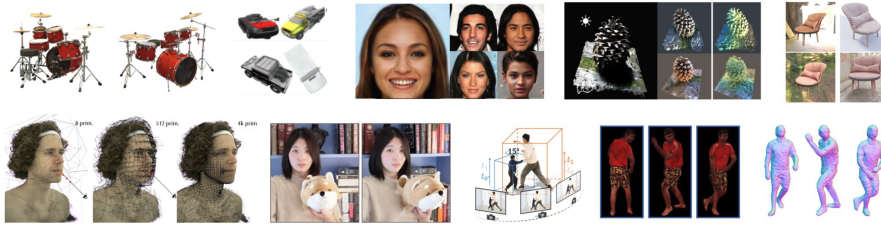


Figure 2.4: Neural rendering encompasses a broad class of techniques for tasks such as novel-view synthesis of static and dynamic scenes, generative object modeling, and photometric scene relighting. Its popularity in digital avatar creation stems from its ability to learn photorealistic human appearance models directly from data. The illustration is courtesy of Tewari *et al.* [177, 178].

Most face representation and tracking pipelines rely on parametric 3D morphable models (3DMMs) [13, 95], which provide a compact and expressive prior for human facial geometry and appearance. In recent years, photorealistic 3D avatar generation has shifted toward neural implicit representations, in particular neural radiance fields (NeRFs) [117] and volumetric primitives such as 3D Gaussians [80], which allow for more faithful modeling of complex appearance and lighting phenomena.

A foundational contribution to NeRF-based avatar modeling is NeRF-Face [44], which integrates a parametric 3D morphable model (3DMM) with a neural radiance field by conditioning the radiance field on expression parameters obtained from the Basel Face Model (BFM) [13, 182]. This paradigm inspired a broad range of subsequent methods [46,

58, 139, 210, 223, 227, 228, 234], which aim to more tightly couple radiance fields to 3DMM geometry. A common strategy is to use the deformation fields defined by the 3DMM to warp volumetric features into canonical space, enabling consistent correspondence across expressions and viewpoints. To further improve visual fidelity, several approaches integrate StyleGAN2-like generative architectures [78] into the NeRF rendering framework. For example, GANAvatar [73], PanoHead [1], and EG3D [21] leverage triplane feature representations to construct high-resolution NeRFs with strong identity preservation and photo-realism. Among them, GANAvatar [73] demonstrates the feasibility of creating personalized avatars from sparse inputs by leveraging the generative prior for texture synthesis and identity consistency. A method closely related to ours is StyleAvatar [191], which employs a 3DMM-based tracking pipeline as a foundation. The method learns a personalized facial avatar using a StyleUNet decoder conditioned on features from a pretrained StyleGAN [78], enabling real-time rendering. Despite its efficiency, StyleAvatar relies heavily on image-to-image translation networks, which can introduce visible artifacts and are sensitive to tracking misalignments. In contrast, our method avoids such issues by employing 3D Gaussian splatting and a learned corrective field, which compensates for tracking inaccuracies directly in 3D space while maintaining geometric consistency and photorealism. IMAvatar [227], another related method inspired by the 3DMM paradigm, learns an implicit representation of appearance jointly with expression-dependent blendshapes and blend skinning weights. The method optimizes an implicit surface representation by combining the ray marching technique of Yariv *et al.* [215] with root-finding of the occupancy function, as introduced in SNARF [26], to compute canonical correspondences of deformed surface points. However, we observed that training IMAvatar is computationally expensive (requiring approximately five days) and can be unstable, with occasional divergence during optimization. Recent approaches have replaced NeRF with more versatile 3D Gaussian Splatting (3DGS) representations [80]. GPHM [209] employs a cascade of MLPs to generate Gaussian primitives anchored to a parametric model, enabling expression control and inversion, albeit conditioning solely on the avatar’s shape. Its successor, GPHMv2 [209], augments this framework with a dynamic module and an expanded dataset, further improving reenactment fidelity. HeadGAP [225] likewise leverages MLPs, incorporating part-based features and additional color conditioning to enhance visual quality. In contrast, SynShot [235] explicitly learns Gaussian primitive parameters by modeling their distribution via a VQ-VAE [189], thereby dispensing with a mesh scaffold at test time by embedding shape information directly in its latent space.



## 2.3 FULL-BODY AVATARS

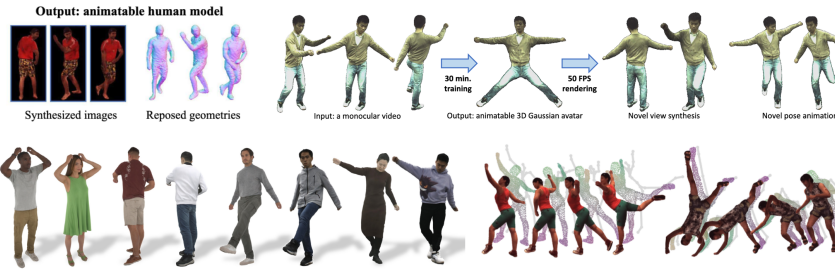


Figure 2.5: Modeling full-body avatars is a highly challenging task, as it involves capturing complex aspects such as body motion, garments, hair, and facial expressions. The figure is courtesy of [65, 98, 132, 172].

Controllable full-body human avatar reconstruction remains a significant challenge in computer vision and computer graphics, as illustrated in Figure 2.5. Existing methods typically rely on dynamic Neural Radiance Fields (NeRF) [116, 128, 129], point-based representations [112, 207, 228], or hybrid approaches [4, 39, 106, 234]. However, these techniques often face limitations such as slow rendering speeds and insufficient disentanglement between garments and body geometry, which restricts their generalization to unseen poses. Recently, the integration of 3D Gaussian Splatting (3DGS) into dynamic human modeling has opened up promising new directions [98, 142, 202, 208, 228]. A notable example is D3GA [231], which enables pose-controllable full-body avatars by leveraging multi-view video sequences and skeletal motion data. This is achieved by combining 3DGS [80] with cage-based deformation techniques [67, 70, 72], allowing efficient animation of dynamic avatars with high visual fidelity.

Neural Radiance Fields (NeRF) [117] have become a de facto standard for avatar appearance modeling, representing scenes as continuous volumetric functions of density and color encoded by a multi-layer perceptron (MLP). Rendering is accomplished via ray marching and volumetric integration of samples along each ray [74]. Numerous works have adapted NeRF to dynamic human performance capture—e.g., HyperNeRF [129], NeuralSF [99], Nerfies [128], D-NeRF [44], Instant-NGP [234], DiNeR [139], PointNeRF [207], and Uni-Warp [212]—achieving high-fidelity animated avatars. However, the majority of these approaches treat the avatar as a single homogeneous volumetric layer [94, 116, 133, 171–173, 229], which limits their ability to model complex phenomena such as loose garments or cloth sliding. To mitigate this, hybrid frameworks [40, 41] fuse explicit parametric geometry (e.g., SMPL [107]) with implicit dynamic NeRFs, improving garment fidelity at the cost of pose generalization. More recently, TECA [218] has extended the SCARF architecture into a generative paradigm, en-

abling text-driven synthesis of NeRF-based accessories and hairstyles via natural language prompts. 3D Gaussian Splatting (3DGS) has recently emerged as a real-time alternative to NeRF, offering interactive frame rates and high-fidelity rendering. Its efficiency and versatility have spurred numerous extensions across diverse domains, including physics-based simulation [203, 226], virtual reality [71], hair modeling [110], head avatar capture [142, 202], fluid dynamics [208], and large-scale scene synthesis [224, 232, 235]. More recently, convolutional regressors have been introduced to predict Gaussian parameters directly from images [98, 126, 151], achieving state-of-the-art visual quality. However, these fixed CNN architectures lack mechanisms for local conditioning or dynamic adjustment of Gaussian counts, and impose a substantial memory overhead—up to 1 GiB due to a two orders of magnitude increase in parameters, which degrades throughput to around 10 FPS [98]. In contrast, the D3GA pipeline remains lightweight and extensible, supporting garment-level decomposition, spatially localized conditioning, and sustained real-time performance.

### 2.3.1 Point-based Rendering

Prior to the emergence of 3D Gaussian Splatting (3DGS), many reconstruction methods employed point-based rendering [112, 172, 228] or sphere-based splatting [90], optimizing both the spatial positions and radii of the primitives during training. For example, NPC [172] defines a point-based body model for avatar representation; however, its reliance on nearest-neighbor searches during training incurs runtimes of up to 12 hours, compared to just 30 minutes for our approach, making it unsuitable for dense multi-view datasets. Ma *et al.* [112] treat garments as a pose-dependent mapping from SMPL body points [107] into a dedicated clothing space, a formulation later enhanced by Prokudin *et al.* [140] through the introduction of a neural deformation field. While both techniques excel at reconstructing garment geometry, neither captures surface appearance. Zheng *et al.* [228] represent the avatar’s upper body with an adaptive point cloud that grows during optimization and is rasterized using a differentiable renderer [195]. Although this method achieves photorealistic local detail, it frequently exhibits artifacts, such as visible holes, that limit its robustness for complete avatar reconstruction.

### 2.3.2 Cage-based Deformations

Cages [124] are a common tool in geometry modeling and animation, acting as sparse proxy structures whose node manipulations propagate deformations to all interior vertices. This yields both effi-

cient computation and intuitive control over complex shapes. Wang *et al.* [194] introduced “neural cages,” in which a learned network rigs a source mesh to a target configuration via a deformable proxy, preserving fine-grained details. Garbin *et al.* [48] extended dynamic NeRF by embedding tetrahedral cages into the volumetric field, allowing ray samples to be “unposed” through tetrahedron–ray intersections. Although this approach delivers real-time performance, high visual fidelity, and precise control, it is best suited for objects with predominantly local deformations (e.g., facial geometry) and does not generalize well to highly articulated, full-body avatars. Peng *et al.* [135] proposed CageNeRF, which applies a low-resolution cage to deform a radiance field for avatar modeling. While their method can be scaled to full-body reconstructions, the coarse cage fails to capture intricate details such as facial features or complex non-rigid motions.

### 2.3.3 Playback Volumetric Videos

Playback methods [17, 38, 69, 97, 199, 214] represent a scene as a time-conditioned function that cannot be arbitrarily controlled, allowing only for a novel viewpoint synthesis while traversing the time axis. Yang *et al.* [214] extended the representation of 3DGS [80] into 4DGS, effectively incorporating time into the primitive representation. Wu *et al.* [199] combine Gaussians with 4D neural voxels, inspired by HexPlane [17], which achieves real-time rendering and novel-view synthesis. However, these solutions fall into a different class of algorithms compared to pose-conditioned drivable avatars, which is our goal.

## 2.4 GENERATIVE MODELING OF HUMANS

StyleGAN [77] initiated a new paradigm in digital human synthesis by enabling the generation of highly realistic human faces. Subsequent work [76, 78] further refined the architecture and improved image fidelity. An alternative generative modeling framework is the variational autoencoder (VAE) [82], which gave rise to volumetric representations such as Neural Volumes [105], the Mixture of Volumetric Primitives method [106], and the broader line of research on codec avatars [4, 114, 115, 144, 151]. More recently, diffusion-based models [63, 123, 148, 165–167] have emerged, outperforming GANs in image synthesis quality. Finally, flow-matching approaches [61, 101, 102, 104] offer yet another powerful paradigm, further advancing avatar generation. In this section, we review these key generative modeling techniques in the context of digital avatars.

### 2.4.1 GAN

Generative adversarial networks (GANs) [56] are a class of generative models that train two neural networks in opposition: a generator, which synthesizes data samples from random noise, and a discriminator, which distinguishes real data from generated samples. During training, the generator improves by attempting to fool the discriminator, while the discriminator becomes more adept at identifying generated outputs (Equation 2.6). This adversarial process drives the generator to produce increasingly realistic data, making GANs particularly effective for high-fidelity image synthesis, style transfer, and other creative generation tasks.

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log(1 - D(G(z)))]. \quad (2.6)$$

StyleGAN has emerged as a versatile backbone for digital avatar synthesis, employed both as a high-fidelity texture generator [3, 91] and as the foundation for full 3D model generation [1, 21, 22, 73]. Pi-GAN [22] unifies color and geometry synthesis by integrating volumetric rendering with a StyleGAN mapping network augmented via FiLM conditioning [137]. EG3 [21] further enhances multi-view consistency by representing features as a triplanar grid: feature maps are predicted on three orthogonal planes, from which volume density and opacity are sampled and subsequently rendered. LatentAvatar [210] incorporates image-based conditioning to guide the generative process, improving fidelity to input poses and expressions. More recently, PanoHead [1] replaces the triplanar representation with a multi-layered voxel grid to eliminate rear-head artifacts and support full-head avatar reconstruction. Despite these advances, common artifacts, particularly around the teeth, remain an open challenge for photorealistic avatar generation. Athar *et al.* [3] propose leveraging a GAN-based prior to enhance in-the-wild facial textures by inpainting missing regions with a model trained on multi-view studio-captured texture data. When combined with the authentic avatars creation framework of Cao *et al.* [18], this approach yields realistic 3D avatars from a single phone scan.

### 2.4.2 VAE

Variational autoencoders (VAEs) [82] are a class of generative models that learn a continuous latent representation of data by combining an encoder network, which maps observations to a parameterized distribution over latent variables, with a decoder network, which reconstructs observations from samples drawn from this distribution. Training proceeds by optimizing a tractable lower bound on the data

log-likelihood, known as the evidence lower bound (ELBO), which balances reconstruction accuracy with regularization of the latent space via a prior distribution (Equation 2.7). This framework enables efficient inference and sampling, making VAEs a versatile tool for tasks such as data generation, representation learning, and unsupervised feature extraction.

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{z \sim q_\phi(z|x)} [\log p_\theta(x | z)] - \text{KL}(q_\phi(z | x) \| p(z)). \quad (2.7)$$

Neural Volumes by Lombardi *et al.* [105] was among the first approaches to model dynamic scenes using a volumetric representation learned via a variational autoencoder (VAE). This work initiated a research direction now known as codec avatars. Ma *et al.* [114] extended this framework by introducing a rendering-adaptive per-pixel decoder, achieving compact and efficient rendering. The Mixture of Volumetric Primitives (MVP) model [106] employs a VAE to infer textures for a  $32^3$  voxel grid, which is subsequently mapped onto a mesh and rendered through volumetric integration. Cao *et al.* [18] further developed a powerful encoder–decoder architecture to construct a prior for avatars derived from a short mobile phone video. Saito *et al.* [151] proposed a VAE-based method that regresses relightable Gaussian primitives instead of voxels [106], enabling ultra-realistic avatar synthesis. Finally, SynShot [235] builds a VAE with latent-code quantization to create a synthetic prior, which is then used for few-shot inversion with pivotal fine-tuning to bridge the domain gap. The compression capabilities of VAEs serve as a backbone for many downstream applications. Esser *et al.* [37] introduce a layered VQ-VAE framework that learns highly expressive codebooks over data distributions, which are subsequently modeled using an autoregressive transformer architecture to produce high-quality images for image synthesis tasks. Rombach *et al.* [148] leverage VAEs to generate latent codes as compressed representations, over which the diffusion process operates, contrasting with full-resolution image-based diffusion in pixel space. For video generation, 3D-VAEs have also proven effective, serving as a video compression method into a temporal latent space. Ho *et al.* [64] demonstrate how a 3D-UNet [29]—which, in essence, extends the original 2D-UNet architecture along the temporal axis and incorporates temporal attention for improved stability—enables generative video diffusion models. This approach has since become widely adopted in the video generation community [15].

### 2.4.3 Diffusion

Diffusion models are likelihood-based generative frameworks that gradually corrupt data through a fixed, forward Markovian diffusion

process and recover it via a learned, reverse denoising process. In the forward diffusion, Gaussian noise is injected over  $T$  small steps as

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I),$$

while the reverse process is modeled by

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)).$$

Training minimizes a simplified denoising objective that provides a tight bound on the negative log-likelihood and produces high-fidelity samples as  $T$  grows:

$$L_{\text{simple}}(\theta) = \mathbb{E}_{t, x_0, \epsilon} \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\|^2, \quad (2.8)$$

where  $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$  and  $\epsilon \sim \mathcal{N}(0, I)$ . For a comprehensive treatment of diffusion-based generative modeling, we refer the reader to the foundational non-equilibrium thermodynamics framework of Sohl-Dickstein *et al.* [165], the score-based perspective of Song and Ermon [167], the original denoising diffusion probabilistic models of Ho *et al.* [63] along with its improvements by Nichol and Dhariwal [123], the development of implicit sampling via DDIM by Song *et al.* [166], and the extension to high-resolution latent spaces by Rombach *et al.* [148]. Diffusion-based methods have spurred extensive research in digital avatar synthesis. GGHead [84] applies a diffusion process over Gaussian primitives to generate static human head representations. Kirschstein *et al.* [83] combine deferred neural rendering with an underlying neural parametric head model to translate geometric cues into photorealistic outputs. Cap4D [175] employs a multiview diffusion model based on Cat3D [45] to learn a dynamic prior from multi-view data, which is then distilled into a real-time 4D avatar. FaceLift [111], following GS-LRM [219], uses a multi-view latent diffusion model (LDM) to regress image-space Gaussian primitives for static faces. Avat3r [86] extends this same concept to dynamic avatars. Recent advancements incorporating Vision Transformers (ViT) [34] and Diffusion Transformers (DiT) [131] have further enhanced representational fidelity. Pippo [75] constructs a prior model leveraging DiT for single-shot inversion of full-body avatars, while Giebenhain *et al.* [55] utilize DiT for tracking and reconstruction via predicted normal maps and optimization-based tracking. Diffusion models have also become increasingly popular for motion generation [193]. Guзов *et al.* [62] employ a head-mounted device to simultaneously generate and reconstruct full-body human motion. Karunratanakul *et al.* [79] leverage pre-trained motion diffusion models as priors for diverse tasks by backpropagating gradients from task-specific criteria defined in motion space through the entire denoising process to refine the latent noise.

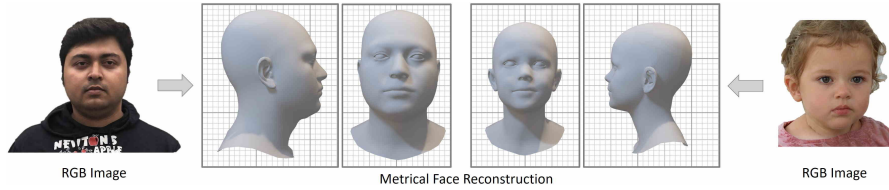


## PUBLICATIONS AND CONTRIBUTIONS

## CONTENTS

3.1	Metrical Accurate Human Shape Regression . . . . .	21
3.2	Instant Head Avatar Creation . . . . .	25
3.3	Full-body 3D Gaussian Avatars . . . . .	28
3.4	Distillation of Avatars into a Linear Model . . . . .	31
3.5	Building a Synthetic Prior for Few-Shot Inversion . . . . .	34

## 3.1 METRICAL ACCURATE HUMAN SHAPE REGRESSION



*Towards Metrical Reconstruction of Human Faces*

Wojciech Zielonka, Timo Bolkart, Justus Thies

*European Conference on Computer Vision (ECCV), Tel-Aviv, Israel, 2022*

## 3.1.1 Motivation

Inferring 3D geometry from 2D images is a fundamentally ill-posed inverse problem [6]. A common approach is to employ a statistical prior such as FLAME [95], which extends the 3D morphable model introduced by Blanz and Vetter in 1999 [13]. However, scale ambiguity persists: under perspective projection, a small face close to the camera can produce the same image as a larger face farther away. Formally, let  $\mathbf{x} \in \mathbb{R}^3$  be a point on the face and  $\mathbf{p} \in \mathbb{R}^2$  its projection onto the image plane. Then

$$\mathbf{p} = \pi(\mathbf{R}\mathbf{x} + \mathbf{t}) = \pi(s[\mathbf{R}\mathbf{x} + \mathbf{t}]) = \pi(\mathbf{R}(s\mathbf{x}) + s\mathbf{t}), \quad (3.1)$$

where  $\pi(\cdot)$  denotes the perspective projection,  $\mathbf{R} \in \mathbb{R}^{3 \times 3}$  the rotation,  $\mathbf{t} \in \mathbb{R}^3$  the translation, and  $s \in \mathbb{R}$  an arbitrary scale factor. Since scaling  $\mathbf{x}$  and  $\mathbf{t}$  by  $s$  leaves  $\mathbf{p}$  unchanged, the true metric size of the

face cannot be recovered without additional information. As a result, these reconstruction techniques can achieve precise 2D alignment yet still fail to recover the true 3D dimensions and spatial placement of the face [33, 39]. However, metrically accurate 3D geometry is crucial whenever the face must interact within a real-world or virtual metric environment. For example, placing the reconstructed face into a virtual reality setting or using it in augmented reality applications, such as AR/VR teleconferencing or virtual try-on, causes these methods to break down, since they do not recover the face’s true scale and shape. To address this, we leverage a face recognition network pretrained on a large 2D dataset [31] to extract robust, identity-discriminative features, and map these features to FLAME shape coefficients by training a regressor network. The regressor is supervised using a high-quality 3D dataset containing mesh-image pairs, thereby creating a 2D-to-3D mapping. Our method inherits the robustness of the recognition network and significantly outperforms prior reconstruction approaches, reducing the average error by 15% on standard benchmarks and by 24% on our proposed evaluation metrics.

### 3.1.2 Results

To comprehensively evaluate our method, we follow two established non-metric benchmarks as well as our proposed metric benchmark. Face shape estimation is assessed on datasets that include reference 3D scans of the subjects. Specifically, we compare against the non-metric NoW Challenge [155] and the benchmark of Feng et al. [42], both of which are used by recent state-of-the-art methods [33, 39, 155], and we additionally report results on our new metric evaluation.

#### 3.1.2.1 Non-Metrical Benchmark

Current evaluation protocols on these datasets incorporate an optimal scaling step when aligning predicted shapes to ground-truth scans, solving for both a rigid transformation and a scale factor, thereby reporting a non-metric (relative) error. This post-hoc scaling masks true shape inaccuracies: for example, on the NoW Challenge [155], the mean error of the average FLAME mesh [95] drops from 1.92 mm to 1.53 mm after scale optimization, a  $\sim 20\%$  reduction despite no actual improvement in reconstruction quality. As a result, these benchmarks can overstate real performance. We evaluate our method under the same protocols and show that it significantly outperforms all prior state-of-the-art approaches.



### 3.1.2.2 *Metrical Benchmark*

Many real-world applications require reconstructions with a true metric scale. To meet this demand, we propose a new evaluation protocol that uses only rigid alignment and forgoes any scale optimization. By enforcing a purely rigid fit, our protocol directly measures metric error, making it fundamentally different from existing relative-error schemes. Additionally, the benchmark by Feng et al. [42] relies on sparse, hand-selected facial landmarks for alignment, a process our experiments show to be highly marker-dependent and prone to inconsistent results. Instead, we adopt the standardized landmark correspondences provided with the FLAME model [95]. Furthermore, we re-evaluate the Feng et al. benchmark using the dense Iterative Closest Point (ICP) method from the NoW Challenge. Across all metrics, our approach delivers substantial improvements in reconstruction accuracy.

### 3.1.2.3 *Metrical Face Tracking*

As an example application, we use our metrically accurate face shape predictions to initialize an analysis-by-synthesis [13] facial expression tracker. Unlike approaches such as [33, 39], our method employs a full perspective camera model, allowing us to recover absolute depth. We measure dense photometric error (RMSE) and observe that our initialization yields substantially lower error than regression-based methods [33, 39]. Moreover, when compared to Face2Face [182], which also uses a perspective model but reports an average depth RMSE of 11.0 mm, our metric face shape estimator reduces this error to just 5.7 mm, significantly enhancing tracking quality.

### 3.1.3 *Discussion*

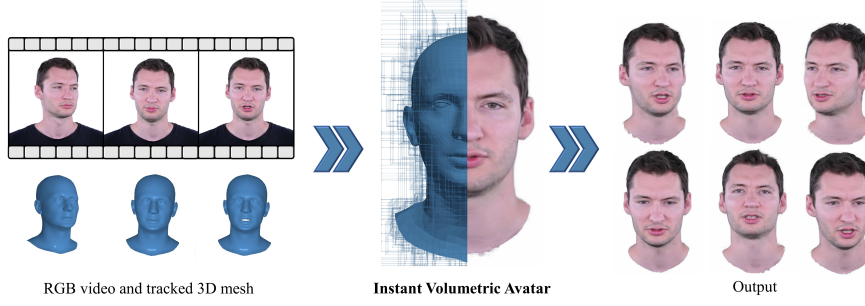
In this work, we evaluated several face-reconstruction methods with respect to their ability to recover metrically accurate 3D shapes. True metric reconstruction is crucial whenever precise measurements of distances and dimensions are required, especially when integrating reconstructed humans into scenes containing objects of known size, such as in virtual or augmented reality applications. We demonstrate that current methods and evaluation protocols are not designed for this purpose. Although existing benchmarks report errors in millimeters, they rely on an optimal scaling step to align the prediction with the ground truth, yielding relative rather than absolute measurements. We contend that this practice is misleading and fails to reflect true metric accuracy. To address this, we introduce a straightforward yet essential modification: eliminate the scale optimization and permit

only rigid alignment between the predicted and reference shapes, enabling genuine metric evaluation.

#### 3.1.4 *Contributions*

To pave the way for metrically accurate reconstructions, we first consolidated existing small- and medium-scale datasets of paired 2D images and 3D scans. This unified dataset enables us to impose direct 3D supervision in our novel shape-prediction framework. Although our combined dataset remains modest in size ( $\sim 2,000$  identities), our model architecture leverages features from a face-recognition network pretrained on a large-scale 2D image corpus, allowing it to generalize to in-the-wild images. Through extensive experiments, we demonstrate state-of-the-art performance on both our newly proposed metric benchmarks and on traditional scale-invariant evaluations. We hope this work encourages the community to focus on metrically grounded face reconstruction and highlights the potential pitfalls of relying solely on non-metric evaluation protocols.

### 3.2 INSTANT HEAD AVATAR CREATION



#### *Instant Volumetric Head Avatars*

Wojciech Zielonka, Timo Bolkart, Justus Thies

*IEEE/CVF Conference on Computer Vision and Pattern Recognition  
(CVPR), Vancouver, Canada, 2023*

#### 3.2.1 Motivation

For immersive AR/VR telepresence, we need digital avatars that not only replicate users’ motion and facial expressions in real time but also match their true shape and appearance. Rather than relying on pre-trained avatars, we propose Instant Volumetric Head Avatars (INSTA), a system that constructs a metrically accurate human avatar in just a few minutes ( $\sim 10$  min) and drives it at interactive frame rates using only commodity hardware and a single RGB camera, where previous techniques require days to train, up to a week in some cases [44, 58, 227]. INSTA leverages dynamic Neural Radiance Fields [44] built on Neural Graphics Primitives [119], embedded around a parametric FLAME face model [95], to achieve fast training and real-time rendering. Crucially, we employ a metrically accurate face reconstruction [233] so avatars have true-to-scale dimensions suitable for environments with known-size objects. We define a canonical space for our Radiance Field and use the FLAME-based motion to drive a deformation field implemented via deformation gradient and bounding volume hierarchy to map points from each frame’s deformed space back into the canonical frame for NeRF evaluation. To capture fine-scale details (e.g., wrinkles, mouth interior), we condition the NeRF on expression parameters. Finally, during training, we further regularize novel-view synthesis by rendering FLAME-derived depth maps as a geometric prior for the NeRF [117].

#### 3.2.2 Results

To evaluate image quality and novel-view extrapolation, we compare our method against NeRFace [44], IMAvatar [227], and Neural Head

Method	L2 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Time (s) ↓
NHA [58]	0.0022	27.71	<b>0.95</b>	<b>0.04</b>	0.63
IMAvatar [227]	0.0023	27.62	0.94	0.06	12.34
NeRFace [44]	<b>0.0018</b>	<b>29.28</b>	<b>0.95</b>	0.07	9.68
Ours	<b>0.0018</b>	28.97	<b>0.95</b>	0.05	<b>0.05</b>

Table 3.1: Average photometric errors over 19 videos spanning our dataset and the public NHA, IMAvatar, and NeRFace benchmarks. “Time” denotes the average rendering time per frame. Our method matches NeRFace on pixel-wise metrics, achieves low perceptual error, and is substantially faster to train and evaluate.

Avatars (NHA) [58]. Quantitatively, we assess photometric accuracy using mean squared error (L2), PSNR, SSIM, and the perceptual metric LPIPS. Note that IMAvatar is trained at a resolution of  $256^2$  due to its computational complexity; for fair comparison, we upsample its outputs to  $512^2$ . All methods produce sharp, photo-realistic frames that closely resemble the ground truth. However, NHA exhibits the most noticeable artifacts, particularly around the ears. IMAvatar suffers from convergence and stability issues on some sequences, leading to optimization failures and premature training termination. In contrast, our approach delivers the best overall image quality while significantly reducing training and inference time. Novel-view extrapolation is critical for 3D avatars in AR/VR applications. We observe that NeRFace produces blurry results around the eyes and teeth, IMAvatar shows silhouette artifacts at grazing angles, and NHA suffers from degraded geometry with strong ear artifacts. By comparison, our method robustly generates photo-realistic images under unseen poses and maintains high fidelity, especially in the skin and mouth regions. In summary, INSTA is orders of magnitude faster than existing state-of-the-art methods while delivering equal or superior avatar quality. Its real-time performance and lightweight design support a wider range of downstream applications, with on-the-fly refinement as new video frames are received.

### 3.2.3 Discussion

Although INSTA outperforms current RGB-video-based avatar methods in both quality and speed, several challenges remain for future work. First, while our model captures dynamic facial expressions, it does not account for changing hair geometry, so hair detail lags behind the fidelity of the face. Second, the employed 3DMM omits teeth geometry; incorporating a more accurate mouth model would improve view extrapolation and render a higher-quality mouth inte-

rior. Third, although we achieve real-time rendering at  $512^2$  resolution, further speed optimizations are needed to support high-resolution AR/VR video conferencing. Finally, with additional engineering, the training stage could run as a background process, continuously refining the canonical avatar after an initial warm-up period and filling in previously unseen regions as they become visible during the session.

#### 3.2.4 Contributions

Instant Volumetric Head Avatars introduces a method for rapidly building fully metric 3D head models from a single RGB video. By embedding Neural Graphics Primitives [119] around a 3D morphable face model [95], we optimize a subject’s dynamic Neural Radiance Field [117] in under ten minutes rather than hours or days. Our approach incorporates a surface embedded radiance field for fast, metrically accurate avatar reconstruction and adds a 3DMM-driven regularization of the density field to improve pose extrapolation, which is critical for AR/VR. Through comparative evaluations and ablation studies, we show that INSTA can generate on-the-fly avatars that match a person’s current appearance, rather than relying on out-of-date, prerecorded models. We believe this shift toward adaptable, real-time avatar creation marks a key advance for immersive telepresence.

### 3.3 FULL-BODY 3D GAUSSIAN AVATARS



*Drivable 3D Gaussian Avatars*

Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, Javier Romero

*International Conference on 3D Vision (3DV), Singapore, Republic of Singapore, 2025*

#### 3.3.1 Motivation

Developing drivable, photorealistic human avatars is key to creating truly immersive long-distance telecommunication experiences. Because facial expressions, body movements, and clothing each follow different deformation patterns, a single-layer model struggles to capture all the nuances. Instead, multi-layered avatars are needed: one layer for the underlying body, another for garments (to handle sliding and folds), and so on. Mixture of Volumetric Primitives (MVP) [106] pioneered a hybrid approach by embedding volumetric elements onto a tracked mesh’s surface, yielding excellent results. Yet it fails if the base mesh is imprecise or lacks detail, causing artifacts and misaligned primitives. Likewise, CNN-based methods [4, 98, 106, 144] fix the number of primitives at training time and offer no straightforward way to decompose garments into separate layers. Moreover, many state-of-the-art techniques [4, 94, 106, 172] cannot condition different parts of the avatar independently. Yet such layered conditioning is critical for realistic motion. Finally, while drivable NeRFs and 3D Gaussian Splatting avatars typically use linear blend skinning (LBS) to map between canonical and posed spaces, LBS’s limited degrees of freedom can’t model complex, non-linear deformations. Tetrahedron-based warping, by contrast, supports richer motion patterns (including stretch) and more physically plausible behavior.

#### 3.3.2 Results

We benchmark D3GA against five state-of-the-art multiview approaches [4, 65, 95, 106, 143]. On our dataset, we compare D3GA to BodyDecoder (BD) [4] and MVP-based avatars [106, 144]. We also evaluate D3GA on the ActorsHQ dataset [69], using just 40 cameras,

alongside Animatable Gaussians (AG) [98], 3DGS-Avatar [143], and Gaussian Avatar (GA) [65], each trained on the same multiview data. Note that D3GA, 3DGS-Avatar, and GA belong to a lightweight class of MLP-based models (up to 10 million parameters). In contrast to the heavy CNN-based MVP, BD, and AG (which require roughly 230 million parameters). All methods are evaluated with SSIM, PSNR, and the perceptual metric LPIPS [220] on random-color backgrounds. For ActorsHQ, we obtain SMPL-X fits via OpenPose [20] and scan-to-mesh optimization. Table 3.2 shows that D3GA achieves the highest PSNR and SSIM on our dataset compared to MVP [106] and BD [4]. On ActorsHQ, D3GA again leads in PSNR and SSIM among Gaussian-based avatars. The slightly reduced sharpness reflects our model’s much smaller size versus the CNN-heavy AG [98]. Furthermore, D3GA supports layer-wise decomposition of the avatar: each garment layer can be driven directly by skeleton joint angles, without additional registration modules.

Dataset	Method	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$
Ours	Ours	<b>30.634</b>	0.054	<b>0.964</b>
	MVP [106]	28.795	0.051	0.955
	BD [4]	29.918	<b>0.044</b>	0.959
ActorsHQ	Ours	<b>26.562</b>	0.065	<b>0.944</b>
	GA [65]	24.731	0.088	0.933
	3DGS-Avatar [143]	21.709	0.082	0.915
	AG [98]	26.454	<b>0.055</b>	0.937

Table 3.2: On our dataset, D3GA achieves the highest PSNR and SSIM compared to BodyDecoder [4] and MVP [106]. Among MLP-based avatars, D3GA also leads in image quality—only Animatable Gaussians (AG), with its larger CNN backbone, produces slightly sharper results.

Our model strikes an effective balance between visual fidelity and parameter count, yielding a compact, easily portable representation. Unlike heavier CNN-based approaches such as AG [98], D3GA matches the footprint of other MLP-based methods while delivering superior image quality. This combination of efficiency and performance makes D3GA especially well-suited for telepresence applications.

### 3.3.3 Discussion

Although D3GA delivers superior visual quality and competitive real-time performance, several challenges remain. Fine, high-frequency

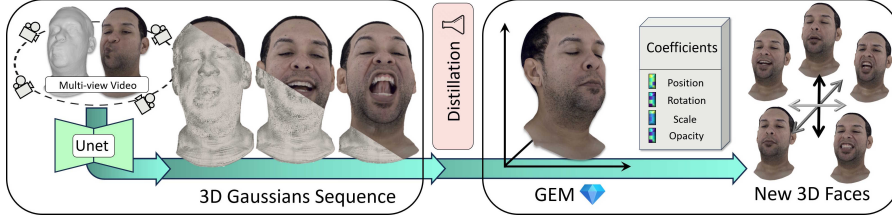
textures (e.g., stripes) can still appear blurred; one remedy might be to regress Gaussian parameters per texel via a variational autoencoder, as in [98, 106]. Despite our regularizers, loose clothing can self-collide or exhibit unrealistic wrinkles and shadows. Adding explicit tetrahedral collision detection could help prevent garment interpenetration. Another promising extension is to swap in the current appearance model for relightable control. Currently, D3GA is demonstrated on a handful of subjects captured in a dense multi-view rig, which both limits real-world deployment and helps prevent misuse (e.g., driving someone’s likeness without permission), but broadening to in-the-wild scenarios is an important future direction. Finally, D3GA’s modular design makes it easy to tailor to specific use cases: one could add more Gaussians for extra detail or drop the garment supervision module when precise cage decomposition isn’t required.

#### 3.3.4 *Contributions*

D3GA introduces a multi-layered framework for animatable human avatars by embedding 3D Gaussian primitives within deformable tetrahedral cages. We warp each Gaussian from the canonical to the posed configuration by directly applying the local deformation gradient to its parameters, resulting in more accurate and artifact-free deformations. Thanks to its compositional design, D3GA supports fine-grained conditioning—using facial keypoints for expressions or other region-specific controls—and can be extended effortlessly to hair, hands, clothing, or accessories. This flexibility is vital for building complete avatars driven by diverse input signals. In our experiments, D3GA delivers higher visual fidelity than comparable state-of-the-art architectures, all while remaining compact, lightweight, and capable of real-time performance.



## 3.4 DISTILLATION OF AVATARS INTO A LINEAR MODEL

*Gaussian Eigen Models for Human Heads*

Wojciech Zielonka, Timo Bolkart, Thabo Beeler, Justus Thies

IEEE/CVF Conference on Computer Vision and Pattern Recognition  
(CVPR), Nashville, USA, 2025

## 3.4.1 Motivation

3D Morphable Models (3DMMs), first introduced by Blanz and Vetter in 1999 [13], perform Principal Component Analysis (PCA) on approximately 200 laser-scanned and registered faces to extract the main modes of variation in geometry and albedo. New faces are generated by specifying values for each principal component, computing a linear combination of those components, and adding the resulting offsets to the mean shape and texture. This linear, mesh-based framework remains the standard for facial performance capture; both regression- and optimization-based, and underpins recent neural rendering-driven 3D avatars [36, 177, 178, 236]. However, traditional 3DMMs often lack fine appearance detail and rely on large CNNs to achieve photorealism. Such models are computationally expensive, slow down rendering, and produce checkpoints exceeding 500 MB, making them impractical for on-device or real-time applications. To overcome these limitations, we introduce GEM (Gaussian Eigen Models for Human Heads), a personalized linear appearance model that uses 3D Gaussians as geometry primitives in the spirit of 3D Gaussian Splatting [80]. Unlike recent Dynamic 3D Gaussian Avatar methods [98, 126, 142, 151, 208, 224, 231], GEM is compact and lightweight, avoiding the need for massive CNN architectures. Our pipeline begins by training a modified U-Net [191] to predict Gaussian parameters in UV space, establishing a consistent representation for each subject. We then collect these Gaussian maps across all training frames and apply PCA to build a small, subject-specific eigenbasis. The final model captures high-fidelity appearance variations with an adaptable number of parameters, enabling efficient distribution and real-time rendering on commodity hardware.

### 3.4.2 Results

We evaluate GEM on the NeRSemble dataset [85], which provides synchronized RGB images from 16 cameras (resolution  $802 \times 550$ ) and corresponding tracked meshes [142]. Our baselines are: Gaussian Avatars (GA) [142], which attaches Gaussians directly to the FLAME mesh without any neural network; Animatable Gaussians (AG) [98], which is a CNN-based Gaussian-map predictor; INSTA [234], which uses dynamic NeRF [117] embedded on the surface of the mesh. Each baseline follows a two-stage pipeline avatar construction and parameter estimation, and often relies on offline tracking with extra objectives (e.g., hair reconstruction [53, 142]), preventing fully real-time operation despite fast rendering. Our approach adds a third stage: building the GEM eigenbasis, which adds only negligible overhead ( $\sim 1$  min) on top of avatar reconstruction. For a fair comparison, we report results for both our full CNN model (**Ours Net**) and the distilled linear model (**Ours GEM**), driven by analysis-by-synthesis fitting [13, 182, 233]. We also include cross-reenactment results using our learned coefficient regressor versus DECA-driven FLAME meshes [39]. All methods are evaluated on novel expressions and novel views using the train/val split from Qian *et al.* [142]. For GEM, we distill 50 PCA components from  $256^2$  Gaussian-texture maps, yielding  $\sim 60k$  active Gaussians. AG uses a similar count for front and back textures, while GA employs around 100k Gaussians. To compute these results, we measure

Method	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	L1 $\downarrow$
AG [98]	29.01	0.0812	0.9429	0.0099
GA [142]	28.31	0.0815	0.9433	0.0102
INSTA [234]	27.92	0.1153	0.9340	0.0128
Ours Net	29.25	0.0777	0.9448	0.0096
Ours GEM	<b>32.68</b>	<b>0.0675</b>	<b>0.9633</b>	<b>0.0069</b>

Table 3.3: Quantitative evaluation on novel expressions and views across 16 cameras. GEM, driven by analysis-by-synthesis fitting, outperforms all baselines in PSNR, LPIPS, SSIM, and L1 error.

image-space color errors: PSNR, LPIPS, L1 loss, and SSIM, following the protocol of Gaussian Avatars [142]. For GEM, we sequentially optimize the PCA coefficients per image using photometric objectives. Unlike the baselines, which require FLAME-based offsets for tracking, GEM can be driven directly, simplifying the overall pipeline. We better capture high-frequency details, pose-dependent wrinkles, and self-shadows, something which is not possible for methods like Gaussian Avatars [142] or INSTA [234], since they either do not use expression-dependent neural networks or limit the conditioning to a small region only.

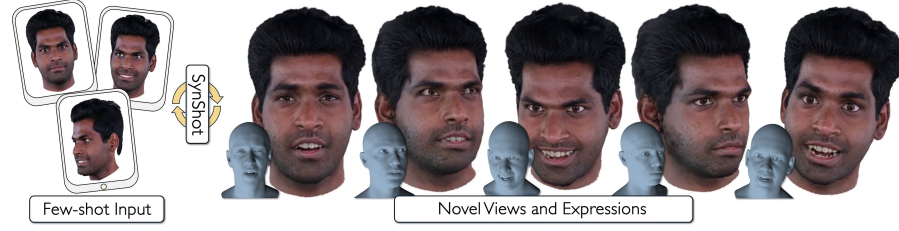
### 3.4.3 Discussion

We propose a universal method to distill any 3D Gaussian Splatting-based avatar system into a compact Gaussian Eigen Model (GEM), assuming the availability of normalized Gaussian-image pairs across the training frames. The only requirement for successful distillation is a dataset that provides these semantically consistent paired maps for all deformation sequences. Our experiments demonstrate that this linear basis representation achieves state-of-the-art performance in both reconstruction quality and runtime speed. To capture fine, wrinkle-level details, the original generator must first produce high-resolution outputs. Furthermore, our distillation pipeline can be applied to existing approaches such as [235], drastically reducing their computational and memory footprint. GEM is well suited to commodity hardware: it synthesizes Gaussian primitives via simple linear combinations of basis vectors, which enables applications in holoportation, audio-driven avatars, and immersive virtual reality. Nevertheless, GEM’s global PCA basis cannot represent very localized deformations or novel feature combinations outside the training set. To address this, future work could introduce a localized PCA decomposition [121], enhancing control over fine-scale variations and broadening the range of expressible motions. Additional limitations include reduced stability in extreme side views and the need to retrain GEM for each new subject using multi-view data. Developing a cross-subject statistical GEM model is an exciting direction to improve generalization.

### 3.4.4 Contributions

We introduce Gaussian Eigen Models for Human Heads (GEM), a linear appearance model for creating photo-realistic head avatars. Its simple formulation dramatically reduces computational cost compared to CNN-based approaches, while still supporting a wide range of applications. The compact representation facilitates easier storage, sharing, and deployment of personalized avatars. GEM also enables real-time avatar animation from RGB input by allowing control over the number of eigenbases, offering a tunable trade-off between memory footprint and visual fidelity. Furthermore, our distillation pipeline can compress existing avatar frameworks into GEM, making them lightweight and resource-efficient. Finally, we demonstrate GEM’s versatility through real-time self-reenactment and cross-person animation scenarios, highlighting its potential for interactive telepresence and virtual reality applications.

## 3.5 BUILDING A SYNTHETIC PRIOR FOR FEW-SHOT INVERSION

*Synthetic Prior for Few-Shot Drivable Head Avatar Inversion*

Wojciech Zielonka, Stephan J. Garbin, Alexandros Lattas, George Kopanas, Paulo Gotardo, Thabo Beeler, Justus Thies, Timo Bolkart

*IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, USA, 2025*

## 3.5.1 Motivation

High-fidelity, drivable digital avatars are essential for immersive virtual and mixed reality experiences. Traditional asset-based pipelines for creating photorealistic human heads demand complex capture setups and extensive manual cleanup, making them both time-consuming and costly [162]. Recent advances in learning-based approaches and neural radiance fields [80, 117] have dramatically simplified avatar creation, yielding high-quality neural head avatars with far less manual effort [44, 114, 191]. Progress includes lightweight animation control via 3D Gaussians [142, 202, 232] and training times reduced to minutes [234]. Most methods train on either multi-view datasets [114, 115, 142, 191, 200] or single-view videos [27, 44, 202, 234], often requiring hundreds to thousands of frames. Such large datasets introduce challenges: robustly tracking a coarse head mesh across all frames, typically by fitting a 3D morphable model [13, 95], can be error-prone, and existing personalized avatars often generalize poorly to expressions and viewpoints unseen during training. To reduce data and capture requirements, recent “few-shot” methods reconstruct avatars from one or a handful of images [28]. However, these approaches typically lag behind large-dataset models in rendering fidelity [142, 232]. Some systems improve quality by first learning a multi-identity head prior, requiring large, multi-view datasets, and then fine-tuning to the target subject [209, 225]. Capturing and managing such datasets is expensive, and compliance with privacy regulations (e.g., GDPR) adds further operational burden, since derivatives and trained models must be purged periodically upon participant request. An alternative is to build priors from large in-the-wild collections such as FFHQ [77], which underpins several GAN-based inversion methods [32, 174, 222]. Yet these often introduce artifacts in novel-view synthesis and struggle to preserve identity. Ultimately, a prior’s expressive power depends on the diversity of its training data (ethnicity, age, facial features, expres-

sions), the quality and calibration of the capture hardware (lighting, camera density, frame rate), and the reliability of preprocessing steps (mesh tracking, background masking). Unlike previous approaches that require expensive real-world capture, SynShot trains its head prior entirely on synthetic 3DMM renders, eliminating costly hardware and privacy concerns. Drawing on successes in synthetic-data methods for 3D face regression [161], landmark prediction [198], and few-shot reconstruction [16, 192, 217], we use a large, diverse synthetic dataset. To adapt to real inputs, we fit this prior and then fine-tune with pivotal tuning [147]. From just three images, SynShot produces a photorealistic 3D Gaussian avatar [80] via a UV-space CNN [98, 151, 232], outperforming state-of-the-art monocular methods [163, 202, 234] and GAN-based ones [32, 174, 222].

### 3.5.2 Results

We compare SynShot to two categories of methods: personalized monocular pipelines and general inversion-based approaches. The personalized monocular method: INSTA [234], Flash Avatar [202], and SplattingAvatar [163]—all rely on FLAME meshes [95]. For these, we evaluate on an ensemble of four datasets [44, 58, 227, 234], each processed with the FLAME tracker from Zielonka *et al.* [233]. SplattingAvatar follows Zheng *et al.* [227] using DECA [39] for pose and expression regression; in our implementation, we replace DECA with an in-house regressor of comparable accuracy. To avoid bias, we select training frames according to the Fibonacci sequence  $F_n = \{1, 2, 3, 5, 8, \dots, 987\}$ . All experiments use progressive farthest-point sampling [141] in the 3DMM expression space to choose frames. Self-reenactment performance is measured on the last 600 frames of the INSTA dataset [234] using LPIPS and SSIM. Finally, we perform a thorough evaluation and compare SynShot against the state-of-the-art inversion-based methods Portrait4D [32], Next3D [174], and InvertAvatar [222], demonstrating that SynShot significantly outperforms all of them.

#### 3.5.2.1 Monocular Avatar Cross-Reenactment

Cross-reenactment evaluates generalization to unseen expressions and viewpoints. Although as few as 13 frames can yield strong performance in-distribution, existing monocular methods [44, 58, 163, 202, 234] often produce artifacts when driven by out-of-distribution sequences. In contrast, SynShot requires only three images. Utilizing the synthetic prior with a clear shape-expression disentanglement network, it outperforms state-of-the-art pipelines trained on thousands of frames, demonstrating the value of a robust prior.

### 3.5.2.2 GAN-based Baselines

We benchmark SynShot against three animatable GAN-based avatars. Both SynShot and InvertAvatar [222] use three input images; Portrait4D [32] and Next3D [174] use a single image. Quantitatively, SynShot achieves an LPIPS of 0.0236, compared to 0.0962 for InvertAvatar, 0.0843 for Portrait4D, and 0.2274 for Next3D, confirming that SynShot significantly outperforms all GAN-based baselines. Qualitatively, SynShot preserves identity and remains stable under novel views and expressions, whereas GAN-based methods often introduce artifacts in side views.

### 3.5.3 Discussion

While SynShot significantly outperforms monocular and GAN-based methods, it still faces several limitations. The primary challenge remains bridging the domain gap between synthetic training data and real-world inputs. In our current pipeline, all synthetic subjects share identical teeth geometry and textures, causing inverted avatars' mouth interior details to adhere too closely to the prior and lack individual variation. Likewise, expression-dependent wrinkles are underrepresented in the synthetic dataset, diminishing fine-scale realism in the reconstructed avatars. Additionally, we render all synthetic heads under a single environment map, which limits generalization to diverse lighting conditions. Future work should focus on enriching the synthetic corpus by varying tooth models, adding wrinkle morphologies, and using multiple environment maps to further improve the visual fidelity of the personalized head avatars.

### 3.5.4 Contributions

We introduce SynShot, a novel method for reconstructing personalized 3D Gaussian head avatars from only a few images. SynShot first trains a generative avatar entirely on synthetic data and then uses it as a prior in an inversion pipeline. This pipeline employs a pivotal tuning strategy that effectively bridges the domain gap between synthetic priors and real input images. We demonstrate that our personalized avatars generalize more robustly to unseen expressions and viewpoints than current state-of-the-art head models. Additionally, the reconstructed Gaussian point cloud can be distilled into a lightweight, network-free representation using GEM [232], removing the requirement for high-end GPU hardware.



## DISCUSSION

---

### CONTENTS

4.1	Summary of Contributions . . . . .	37
4.2	Potential Limitations . . . . .	39
4.3	Future Work . . . . .	40
4.4	Conclusions . . . . .	41

---

This thesis focuses on a broad range of topics within the field of digital humans, including the reconstruction of metrically accurate heads, real-time avatar creation, full-body avatars, efficient storage and representation of Gaussian avatars, and few-shot inversion. All of these components are crucial for building a holistic system capable of representing avatars that are indistinguishable from reality. However, this goal remains ahead, as it requires integrating additional aspects not addressed in this work, such as hair and garment modeling, as well as human–scene interactions. With the advent of powerful new representations such as flow matching models and diffusion-based approaches, the field of digital humans is expected to progress rapidly, enabling novel applications such as real-time telepresence, holographic teleportation, and extended VR/MR systems that enhance everyday communication and productivity. Moreover, sparse 3D datasets can be effectively complemented with large-scale 2D datasets, consisting of videos and images, to build strong priors for avatar inversion, scene interaction, motion generation, and many other tasks, fundamentally shifting the paradigms of 3D avatar creation as known so far. In this section, we summarize the contributions made in this work, discuss their limitations, outline directions for future research, and conclude with final remarks.

### 4.1 SUMMARY OF CONTRIBUTIONS

Section 3.1 introduced MICA [233], a regression-based method that advances metrically accurate prediction of human head geometry. The pipeline was trained on a combined dataset composed of several smaller datasets, each containing paired 3D geometry and 2D images.

By leveraging the robust face recognition network ArcFace [31], the method effectively maps 2D image features to 3D human shape, supporting the hypothesis that ArcFace implicitly learns geometric information from images alone. As a practical application, we implemented a monocular metrically-accurate face tracker, which outperformed existing methods due to significantly improved initialization of the predicted shape using MICA.

Section 3.2 introduced INSTA [234], a novel method for instantaneously reconstructing an avatar and driving it in real-time. This was achieved by incorporating a deformation field represented by the Jacobian matrix between the canonical and deformed spaces. Leveraging this information, NeRF samples could be efficiently mapped from the deformed space to the canonical space using the deformation gradient of the closest triangle. Real-time performance was enabled through the use of a bounding volume hierarchy (BVH) built around the mesh to facilitate fast nearest-neighbor searches. INSTA was also capable of continuously updating the canonical space based on a stream of input images, allowing for dynamically adjustable quality. This work represents a significant step toward on-demand avatar creation without relying on personalized networks trained over several days, as was the case in previous approaches.

Section 3.3 presents D3GA [231], a novel method with the following contributions: a lightweight, flexible, and composable model based on 3D Gaussian primitives, driven by tetrahedral cage-based deformations that enhance body modeling capabilities; and localized motion conditioning that enables the representation of fine-grained motions such as facial expressions. D3GA was among the first methods to combine Gaussian primitives with a parametric body model. Additionally, by leveraging tetrahedral structures, deformations could be more effectively transferred to the Gaussian kernels, allowing the model to represent complex transformations such as stretching and rotation of the encapsulated shape. The usefulness of this approach was further supported by concurrent works that adopted similar strategies for scene representation using tetrahedralization [57, 60].

Section 3.4 describes GEM [232], a method that distills neural networks into a linear and lightweight 3D Gaussian head avatar, represented as an ensemble of eigenbases. As a demonstration of this representation, we developed a real-time, cross-person animation system that drives GEM avatars from single input images using a generalizable regressor. This approach was designed with efficiency in mind, targeting scenarios where models must be transmitted over the wire. While high-quality neural representations typically require hundreds of megabytes, GEM provides a significantly more compact alternative, orders of magnitude smaller, while still preserving high fidelity and expression-dependent detail. Finally, we showcased a real-time



pipeline capable of transferring expressions between different actors. We believe this representation holds great promise for telepresence applications and enables deployment on commodity devices, as it does not require a high-end GPU for inference.

Section 3.5 showcases SynShot [235], a novel generative method based on a convolutional encoder–decoder architecture trained exclusively on large-scale synthetic data to produce controllable 3D head avatars. In addition, it introduces a reconstruction scheme that adapts and fine-tunes a pretrained generative model using only a few real images to create a personalized, photorealistic 3D head avatar. Given as few as three input images, SynShot can accurately project them onto the VAE’s latent manifold and successfully bridge the domain gap introduced by training solely on synthetic data. This demonstrates that synthetic data can be effectively leveraged to build powerful generative prior models applicable to a wide range of downstream tasks.

## 4.2 POTENTIAL LIMITATIONS

Each of the introduced methods has its limitations. While the field is gradually progressing toward photorealistic digital avatars, several challenges remain unresolved. MICA, for instance, is trained on a relatively small dataset of approximately 2,000 identities, which limits its diversity and generalization capability. This restricts its effectiveness in handling a wide range of facial appearances, particularly under varying lighting or occlusion conditions. INSTA does not model dynamically changing hair, resulting in hair quality that lags behind the fidelity of the facial interior. Improvements in level-of-detail and temporal consistency are needed to close this gap. Moreover, the underlying 3D Morphable Model (3DMM) used in INSTA lacks teeth geometry, which affects the quality of the mouth region, especially when rendering from novel viewpoints. While the method achieves real-time frame rates for rendering at a resolution of  $512^2$ , rendering speed remains a bottleneck for high-resolution. D3GA is currently limited to modeling photorealistic avatars for a small number of consenting subjects captured using a dense multi-view setup. In addition, handling self-collisions for loose garments remains an open challenge, as the sparse control signals fail to convey sufficient information about high-frequency deformations like wrinkles or self-shadowing. The PCA-based GEM models limit the ability to perform fine-grained, localized edits. Incorporating a localized PCA basis could enhance controllability and allow the synthesis of a broader range of expressions beyond those seen during training. Other challenges include limited generalization to side-view imagery, leading to unstable expressions and weak personalization. For each new subject, a new

model must be learned from multi-view observations. SynShot faces challenges stemming from the lack of diversity in its synthetic training dataset. For example, all synthetic subjects share the same teeth geometry and texture, causing the inverted avatars to rigidly follow the prior and limiting personalization. The dataset also fails to capture expression-dependent fine details such as wrinkles, which compromises the realism of the output. Additionally, all scenes were ray-traced under a single environment map, reducing generalization to diverse lighting conditions encountered in real-world environments. In summary, while each method contributes uniquely to the advancement of digital human modeling, addressing their respective limitations is crucial for achieving truly generalizable, high-fidelity, and real-time avatars suitable for practical applications. One of the most significant current challenges in training large-scale models is the lack of sufficiently large and diverse 3D datasets. In this context, leveraging rich 2D priors emerges as a promising direction to supplement sparse 3D data, enabling the development of more robust and scalable solutions.

#### 4.3 FUTURE WORK

As mentioned previously, one of the major challenges in the field of digital humans is the lack of sufficiently large 3D datasets. The most widely used datasets, such as Nersemble [85], AVA-256 [115], and FaceScape [230], contain only a few hundred subjects. This scale is insufficient for training generalizable models capable of representing the vast diversity of human appearances. In contrast, 2D datasets consisting of images and videos, such as LAION-5B [160], contain billions of samples. Future research will need to leverage such large-scale 2D data, either through hybrid approaches or purely 2D-based solutions, to meet the diversity requirements of robust digital human models. The emergence of video diffusion models [5, 14, 64, 100], with increasing improvements in character consistency and camera controllability [59], raises the question of how far these models can be pushed in terms of realism and generalization. Another important consideration in the growing popularity of Video Foundation Models (VFMs) is their computational cost during inference. In contrast, 3D models and neural representations benefit from decades of advancements in efficient rasterization and ray tracing, enabling low-cost rendering of individual frames. VFMs, however, still incur significantly higher inference costs, both in terms of computation and memory, often making them impractical for deployment on commodity devices. To fully harness their potential, further research is needed to improve their efficiency and reduce their hardware requirements.

Another promising direction for future work is the exploration of human-scene and human-human interactions, which are essential

for achieving fully immersive experiences in VR and mixed reality (MR) environments. Currently, many photorealistic full-body avatar systems [4, 98, 144, 231] lack awareness of their surroundings. At the same time, methods for scene understanding [201, 213] and interaction [204, 221] primarily focus on simulating motion and contact dynamics, often in isolation from avatar realism. To enable the joint simulation of realistic human motion [2, 10, 122, 138] alongside accurate scene reconstruction and understanding, novel integrated approaches are required. These approaches should bridge the gap between photorealistic avatar rendering and physically plausible interactions within complex, dynamic environments. Ultimately, progress in the virtual simulation of scenes or reconstructed environments could be leveraged in fields such as robotics, enhancing both mobility and perception from a software-centric perspective.

This point leads to the ultimate objective of developing agentic artificial humans (robots). Pearl’s foundational work [130] introduces the Structural Causal Model (SCM) framework and the do-calculus, which together form the theoretical backbone of formal causal reasoning in AI. This trajectory in the evolution of artificial intelligence suggests that future models will be equipped with agency [89, 149, 158], enabling them to actively explore and learn from their environment in a manner analogous to human behavior. In the context of digital humans, this implies that, given a suitable environment, such agents could interact both with their surroundings and with one another, potentially rendering them indistinguishable from real humans in terms of behavior, adaptability, and ultimately, appearance.

#### 4.4 CONCLUSIONS

Digital avatars represent a highly diverse and complex research domain. This thesis addresses several key challenges, ranging from facial reconstruction and tracking to full-body avatar modeling. Each of these components is essential to the goal of achieving photorealistic digital humans. However, they also pose significant challenges, as humans are particularly sensitive to even subtle artifacts. Moreover, in a world increasingly shaped by artificial intelligence, where progress continues to accelerate, numerous applications for digital avatars are rapidly emerging. The popularity of tools such as ChatGPT highlights the demand for fully embodied conversational agents that are photorealistic and perceptually familiar to the human eye. Beyond conversational AI, the range of applications continues to grow, encompassing fields such as healthcare, aging, autism support, and storytelling. As avatars transition from passive renderings to interactive agents, their role within virtual ecosystems, such as games, simulations, and metaverse environments, will become increasingly central. This expanding

landscape creates a growing need for lifelike, human-centered digital agents that seamlessly integrate multiple modalities, including speech, facial expression, gaze, and gesture, all synchronized in real time. Furthermore, as digital avatars become more lifelike and autonomous, concerns surrounding identity protection, deepfakes, and the ethical use of synthesized humans become increasingly important. Ensuring trust, interpretability, and appropriate safeguards is just as critical as achieving high visual fidelity. In parallel with these ethical challenges, the rapid growth of large-scale AI models also raises concerns about sustainability. Training and deploying large language models (LLMs) require substantial computational resources, leading to significant energy consumption, especially during large-scale inference. This underscores the importance of developing efficient, modular, and adaptive avatar systems that balance realism and performance with computational cost, ultimately enabling broader accessibility and responsible deployment of this technology. In this thesis, we present methods that advance the state of the art in both the geometric capture and photorealistic synthesis of digital humans. Our work addresses key challenges in scalability, realism, and personalization, contributing toward the long-term vision of digital avatars that are not only visually indistinguishable from real humans but also robust and controllable. While substantial progress has been made, realizing the full potential of generalizable and interactive digital humans remains an open challenge, one that will require continued innovation at the intersection of vision, graphics, and learning.



## APPENDIX

---

### CONTENTS

A.1	Towards Metrical Reconstruction of Human Faces . . .	43
A.2	Instant Volumetric Head Avatars . . . . .	66
A.3	Drivable 3D Gaussian Avatars . . . . .	78
A.4	Gaussian Eigen Models for Human Heads . . . . .	91
A.5	Synthetic Prior for Few-Shot Drivable Head Avatar In- version . . . . .	103
A.6	Broader Impact: Ethical Concerns . . . . .	116

---

### A.1 TOWARDS METRICAL RECONSTRUCTION OF HUMAN FACES

*Towards Metrical Reconstruction of Human Faces*

Wojciech Zielonka, Timo Bolkart, Justus Thies

Published in *European Conference on Computer Vision (ECCV)*, Tel-Aviv,  
Israel, 2022.

#### Abstract

Face reconstruction and tracking are building blocks of numerous applications in AR/VR, human-machine interaction, as well as medical applications. Most of these applications rely on a metrically correct prediction of the shape, especially when the reconstructed subject is put into a metrical context (i.e., when there is a reference object of known size). A metrical reconstruction is also needed for any application that measures distances and dimensions of the subject (e.g., to virtually fit a glasses frame). State-of-the-art methods for face reconstruction from a single image are trained on large 2D image datasets in a self-supervised fashion. However, due to the nature of a perspective projection, they are not able to reconstruct the actual face dimensions, and even predicting the average human face outperforms some of these methods in a metrical sense. To learn the actual shape

of a face, we argue for a supervised training scheme. Since there exists no large-scale 3D dataset for this task, we annotated and unified small- and medium-scale databases. The resulting unified dataset is still a medium-scale dataset with more than 2k identities, and training purely on it would lead to overfitting. To this end, we take advantage of a face recognition network pretrained on a large-scale 2D image dataset, which provides distinct features for different faces and is robust to expression, illumination, and camera changes. Using these features, we train our face shape estimator in a supervised fashion, inheriting the robustness and generalization of the face recognition network. Our method, which we call MICA (MetrIC fAce), outperforms the state-of-the-art reconstruction methods by a large margin, both on current non-metric benchmarks as well as on our metric benchmarks (15% and 24% lower average error on NoW, respectively).

# Towards Metrical Reconstruction of Human Faces

Wojciech Zielonka, Timo Bolkart, and Justus Thies

Max Planck Institute for Intelligent Systems, Tübingen

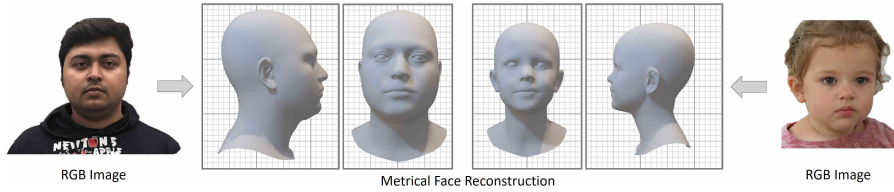


Fig. 1: An RGB image of a subject serves as input to MICA, which predicts a metrical reconstruction of the human face. Images from NoW [59], StyleGan2 [38].

**Abstract.** Face reconstruction and tracking is a building block of numerous applications in AR/VR, human-machine interaction, as well as medical applications. Most of these applications rely on a metrically correct prediction of the shape, especially, when the reconstructed subject is put into a metrical context (i.e., when there is a reference object of known size). A metrical reconstruction is also needed for any application that measures distances and dimensions of the subject (e.g., to virtually fit a glasses frame). State-of-the-art methods for face reconstruction from a single image are trained on large 2D image datasets in a self-supervised fashion. However, due to the nature of a perspective projection they are not able to reconstruct the actual face dimensions, and even predicting the average human face outperforms some of these methods in a metrical sense. To learn the actual shape of a face, we argue for a supervised training scheme. Since there exists no large-scale 3D dataset for this task, we annotated and unified small- and medium-scale databases. The resulting unified dataset is still a medium-scale dataset with more than 2k identities and training purely on it would lead to overfitting. To this end, we take advantage of a face recognition network pretrained on a large-scale 2D image dataset, which provides distinct features for different faces and is robust to expression, illumination, and camera changes. Using these features, we train our face shape estimator in a supervised fashion, inheriting the robustness and generalization of the face recognition network. Our method, which we call MICA (Metric fAce), outperforms the state-of-the-art reconstruction methods by a large margin, both on current non-metric benchmarks as well as on our metric benchmarks (15% and 24% lower average error on NoW, respectively).

**Project website:** <https://zielon.github.io/mica/>

## 1 Introduction

Learning to reconstruct 3D content from 2D imagery is an ill-posed inverse problem [4]. State-of-the-art RGB-based monocular facial reconstruction and tracking methods [18, 23] are based on self-supervised training, exploiting an underlying metrical face model which is constructed using a large-scale dataset of registered 3D scans (e.g., 33000 scans for the FLAME [47] model). However, when assuming a perspective camera, the scale of the face is ambiguous since a large face can be modeled by a small face that is close to the camera or a gigantic face that is far away. Formally, a point  $\mathbf{x} \in \mathbb{R}^3$  of the face is projected to a point  $\mathbf{p} \in \mathbb{R}^2$  on the image plane with the projective function  $\pi(\cdot)$  and a rigid transformation composed of a rotation  $\mathbf{R} \in \mathbb{R}^{3 \times 3}$  and a translation  $\mathbf{t} \in \mathbb{R}^3$ :

$$\mathbf{p} = \pi(\mathbf{R} \cdot \mathbf{x} + \mathbf{t}) = \pi(s \cdot (\mathbf{R} \cdot \mathbf{x} + \mathbf{t})) = \pi(\mathbf{R} \cdot (s \cdot \mathbf{x}) + (s \cdot \mathbf{t})).$$

The perspective projection is invariant to the scaling factor  $s \in \mathbb{R}$ , and thus, if  $\mathbf{x}$  is scaled by  $s$ , the rigid transformation can be adapted such that the point still projects onto the same pixel position  $\mathbf{p}$  by scaling the translation  $\mathbf{t}$  by  $s$ . In consequence, face reconstruction methods might result in a good 2D alignment but can fail to reconstruct the metrical 3D surface and the meaningful metrical location in space. However, a metric 3D reconstruction is needed in any scenario where the face is put into a metric context. E.g., when the reconstructed human is inserted into a virtual reality (VR) application or when the reconstructed geometry is used for augmented reality (AR) applications (teleconferencing in AR/VR, virtual try-on, etc.). In these scenarios, the methods mentioned above fail since they do not reproduce the correct scale and shape of the human face. In the current literature [25, 59, 83], we also observe that methods use evaluation measurements not done in a metrical space. Specifically, to compare a reconstructed face to a reference scan, the estimation is aligned to the scan via Procrustes analysis, including an optimal scaling factor. This scaling factor favors the estimation methods that are not metrical, and the reported numbers in the publications are misleading for real-world applications (relative vs. absolute/metric error). In contrast, we aim for a metrically correct reconstruction and evaluation that directly compares the predicted geometry to the reference data without any scaling applied in a post-processing step which is fundamentally different. As discussed above, the self-supervised methods in the literature do not aim and cannot reconstruct a metrically correct geometry. However, training these methods in a supervised fashion is not possible because of the lack of data (no large-scale 3D dataset is available). Training on a small- or medium-scale 3D dataset will lead to overfitting of the networks (see study in the supplemental document). To this end, we propose a hybrid method that can be trained on a medium-scale 3D dataset, reusing powerful descriptors from a pretrained face recognition network (trained on a large-scale 2D dataset). Specifically, we propose the usage of existing 3D datasets like LYHM [16], FaceWarehouse [10], Stirling [26], etc., that contain RGB imagery and corresponding 3D reconstructions to learn a metrical reconstruction of the human head. To use these 3D datasets, significant work



has been invested to unify the 3D data (i.e., to annotate and non-rigidly fit the FLAME model to the different datasets). This unification provides us with meshes that all share the FLAME topology. Our method predicts the head geometry in a neutral expression, only given a single RGB image of a human subject in any pose or expression. To generalize to unseen in the wild images, we use a state-of-the-art face recognition network [17] that provides a feature descriptor for our geometry-estimating network. This recognition network is robust to head poses, different facial expressions, occlusions, illumination changes, and different focal lengths, thus, being ideal for our task (see Figure 3). Based on this feature, we predict the geometry of the face with neutral expression within the face space spanned by FLAME [47], effectively disentangling shape and expression. As an application, we demonstrate that our metrical face reconstruction estimator can be integrated in a new analysis-by-synthesis face tracking framework which removes the requirement of an identity initialization phase [70]. Given the metrical face shape estimation, the face tracker is able to predict the face motion in a metrical space.

In summary, we have the following contributions:

- a dataset of 3D face reference data for about 2300 subjects, built by unifying existing small- and medium-scale datasets under common FLAME topology.
- a metrical face shape predictor – MICA – which is invariant to expression, pose and illumination, by exploiting generalized identity features from a face recognition network and supervised learning.
- a hybrid face tracker that is based on our (learned) metrical reconstruction of the face shape and an optimization-based facial expression tracking.
- a metrical evaluation protocol and benchmark, including a discussion on the current evaluation practise.

## 2 Related Work

Reconstructing human faces and heads from monocular RGB, RGB-D, or multi-view data is a well-explored field at the intersection of computer vision and computer graphics. Zollhöfer et al. [85] provide an extensive review of reconstruction methods, focusing on optimization-based techniques that follow the principle of analysis-by-synthesis. Primarily, the approaches that are based on monocular inputs are based on a prior of face shape and appearance [6, 7, 27, 28, 40, 66–71, 77, 78]. The seminal work of Blanz et al. [8] introduced such a 3D morphable model (3DMM), which represents the shape and appearance of a human in a compressed, low-dimensional, PCA-based space (which can be interpreted as a decoder with a single linear layer). There is a large corpus of different morphable models [21], but the majority of reconstruction methods use either the Basel Face Model [8, 52] or the Flame head model [47]. Besides using these models for an analysis-by-synthesis approach, there is a series of learned regression-based methods. An overview of these methods is given by Morales et al. [50]. In the following, we will discuss the most relevant related work for monocular RGB-based reconstruction methods.

*Optimization-based Reconstruction of Human Faces.* Along with the introduction of a 3D morphable model for faces, Blanz et al. [8] proposed an optimization-based reconstruction method that is based on the principle of analysis-by-synthesis. While they used a sparse sampling scheme to optimize the color reproduction, Thies et al. [69, 70] introduced a dense color term considering the entire face region that is represented by a morphable model using differentiable rendering. This method has been adapted for avatar digitization from a single image [36] including hair, is used to reconstruct high-fidelity facial reflectance and geometry from a single images [79], for reconstruction and animation of entire upper bodies [71], or avatars with dynamic textures [51]. Recently, these optimization-based methods are combined with learnable components such as surface offsets or view-dependent surface radiance fields [32]. In addition to a photometric reconstruction objective, additional terms based on dense correspondence [35] or normal [1, 32] estimations of neural network can be employed. Optimization-based methods are also used as a building block for neural rendering methods such as deep video portraits [40], deferred neural rendering [68], or neural voice puppetry [67]. Note that differentiable rendering is not only used in neural rendering frameworks but is also a key component for self-supervised learning of regression-based reconstruction methods covered in the following.

*Regression-based Reconstruction of Human Faces.* Learning-based face reconstruction methods can be categorized into supervised and self-supervised approaches. A series of methods are based on synthetic renderings of human faces to perform a supervised training of a regressor that predicts the parameters of a 3D morphable model [20, 41, 56, 57]. Genova et al. [31] propose a 3DMM parameter regression technique that is based on synthetic renderings (where ground truth parameters are available) and real images (where multi-view identity losses are applied). It uses FaceNet [60] to extract features for the 3DMM regression task. Tran et al. [72] and Chang et al. [11] (ExpNet) directly regress 3DMM parameters using a CNN trained on fitted 3DMM data. Tu et al. [75] propose a dual training pass for images with and without 3DMM fittings. Jackson et al. [37] propose a model-free approach that reconstructs a voxel-based representation of the human face and is trained on paired 2D image and 3D scan data. PRN [24] is trained on 'in-the-wild' images with fitted 3DMM reconstructions [84]. It is not restricted to a 3DMM model space and predicts a position map in the UV-space of a template mesh. Instead of working in UV-space, Wei et al. [76] propose to use graph convolutions to regress the coordinates of the vertices. MoFA [65] is a network trained to regress the 3DMM parameters in a self-supervised fashion. As a supervision signal, it uses the dense photometric losses of Face2Face [70]. Within this framework, Tewari et al. proposed to refine the identity shape and appearance [64] as well as the expression basis [63] of a linear 3DMM. In a similar setup, one can also train a non-linear 3DMM [74] or personalized models [12]. RingNet [59] regresses 3DMM parameters and is trained on 2D images using losses on the reproduction of 2D landmarks and shape consistency (different images of the same subject) and shape inconsistency (images of different subjects) losses. DECA [23] extends RingNet with

expression dependent offset predictions in UV space. It uses dense photometric losses to train the 3DMM parameter regression and the offset prediction network. This separation of a coarse 3DMM model and a detailed bump map has been introduced by Tran et al. [73]. Chen et al. [13] use a hybrid training composed of self-supervised and supervised training based on renderings to predict texture and displacement maps. Deng et al. [18] train a 3DMM parameter regressor based on multi-image consistency losses and 'hybrid-level' losses (photometric reconstruction loss with skin attention masks, and a perception-level loss based on FaceNet [60]). On the NoW challenge [59], DECA [23] and the method of Deng et al. [18] show on-par state-of-the-art results. Similar to DECA's offset prediction, there are GAN-based methods that predict detailed color maps [29, 30] or skin properties [44, 45, 58, 79] (e.g., albedo, reflectance, normals) in UV-space of a 3DMM-based face reconstruction. In contrast to these methods, we are interested in reconstructing a metrical 3D representation of a human face and not fine-scale details. Self-supervised methods suffer from the depth-scale ambiguity (the face scale, translation away from the camera, and the perspective projection are ambiguous) and, thus, predict a wrongly scaled face, even though 3DMM models are by construction in a metrical space. We rely on a strong supervision signal to learn the metrical reconstruction of a face using high-quality 3D scan datasets which we unified. In combination with an identity encoder [17] trained on in-the-wild 2D data, including occlusions, different illumination, poses, and expressions, we achieve robust geometry estimations that significantly outperform state-of-the-art methods.

### 3 Metrical Face Shape Prediction

Based on a single input RGB image  $I$ , MICA aims to predict a metrical shape of a human face in a neutral expression. To this end, we leverage both 'in-the-wild' 2D data as well as metric 3D data to train a deep neural network, as shown in Figure 2. We employ a state-of-the-art face recognition network [17] which is trained on 'in-the-wild' data to achieve a robust prediction of an identity code, which is interpreted by a geometry decoder.

*Identity Encoder.* As an identity encoder, we leverage the ArcFace [17] architecture which is pretrained on Glint360K [2]. This ResNet100-based network is trained on 2D image data using an additive angular margin loss to obtain highly discriminative features for face recognition. It is invariant to illumination, expression, rotation, occlusion, and camera parameters which is ideal for a robust shape prediction. We extend the ArcFace architecture by a small mapping network  $\mathcal{M}$  that maps the ArcFace features to our latent space, which can then be interpreted by our geometry decoder:

$$\mathbf{z} = \mathcal{M}(\text{ArcFace}(I)),$$

where  $\mathbf{z} \in \mathbb{R}^{300}$ . Our mapping network  $\mathcal{M}$  consists of three fully-connected linear hidden layers with ReLU activation and the final linear output layer.

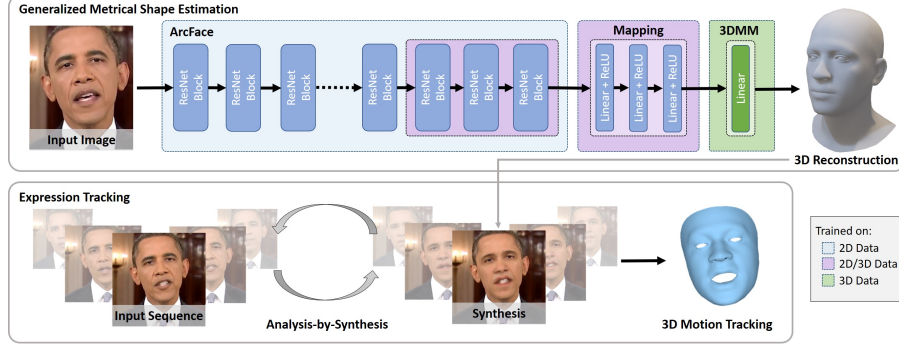


Fig. 2: We propose a method for metrical human face shape estimation from a single image which exploits a supervised training scheme based on a mixture of different 2D, 2D/3D and 3D datasets. This estimation can be used for facial expression tracking using analysis-by-synthesis which optimizes for the camera intrinsics, as well as the per-frame illumination, facial expression and pose.

*Geometry Decoder.* There are essentially two types of geometry decoders used in the literature, model-free and model-based. Throughout the project of this paper, we conducted experiments on both types and found that both perform similarly on the evaluation benchmarks. Since a 3DMM model efficiently represents the face space, we focus on a model-based decoder. Specifically, we use FLAME [47] as a geometry decoder, which consists of a single linear layer:

$$\mathcal{G}_{3DMM}(z) = \mathbf{B} \cdot z + \mathbf{A},$$

where  $\mathbf{A} \in \mathbb{R}^{3N}$  is the geometry of the average human face and  $\mathbf{B} \in \mathbb{R}^{3N \times 300}$  contains the principal components of the 3DMM and  $N = 5023$ .

*Supervised Learning.* The networks described above are trained using paired 2D/3D data from existing, unified datasets  $\mathcal{D}$  (see Section 5). We fix large portions of the pre-trained ArcFace network during the training and refine the last 3 ResNet blocks. Note that ArcFace is trained on a much larger amount of identities, therefore, refining more hidden layers results in worse predictions due to overfitting. We found that using the last 3 ResNet blocks gives the best generalization (see supplemental document). The training loss is:

$$\mathcal{L} = \sum_{(I, \mathcal{G}) \in \mathcal{D}} |\kappa_{mask}(\mathcal{G}_{3DMM}(\mathcal{M}(\text{ArcFace}(I))) - \mathcal{G})|, \quad (1)$$

where  $\mathcal{G}$  is the ground truth mesh and  $\kappa_{mask}$  is a region dependent weight (the face region has weight 150.0, the back of the head 1.0, and eyes with ears 0.01). We use AdamW [49] for optimization with fixed learning rate  $\eta = 1e-5$  and weight decay  $\lambda = 2e-4$ . We select the best performing model based on the validation set loss using the Florence dataset [3]. The model was trained for 160k steps on Nvidia Tesla V100.

## 4 Face Tracking

Based on our shape estimate, we demonstrate optimization-based face tracking on monocular RGB input sequences. To model the non-rigid deformations of the face, we use the linear expression basis vectors and the linear blend skinning of the FLAME [47] model, and use a linear albedo model [22] to reproduce the appearance of a subject in conjunction with a Lambertian material assumption and a light model based on spherical harmonics. We adapt the analysis-by-synthesis scheme of Thies et al. [70]. Instead of using a multi-frame model-based bundling technique to estimate the identity of a subject, we use our one-shot shape identity predictor. We initialize the albedo and spherical harmonics based on the same first frame using the energy:

$$E(\phi) = w_{dense}E_{dense}(\phi) + w_{lmk}E_{lmk}(\phi) + w_{reg}E_{reg}(\phi), \quad (2)$$









where  $\phi$  is the vector of unknown parameters we are optimizing for. The energy terms  $E_{dense}(\phi)$  and  $E_{reg}(\phi)$  measure the dense color reproduction of the face ( $\ell_1$ -norm) and the deviation from the neutral pose respectively. The sparse landmark term  $E_{lmk}(\phi)$  measures the reproduction of 2D landmark positions (based on Google’s mediapipe [33, 39] and Face Alignment [9]). The weights  $w_{dense}$ ,  $w_{lmk}$  and  $w_{reg}$  balance the influence of each sub-objectives on the final loss. In the first frame vector  $\phi$  contains the 3DMM parameters for albedo, expression, and rigid pose, as well as the spherical harmonic coefficients (3 bands) that are used to represent the environmental illumination [54]. After initialization, the albedo parameters are fixed and unchanged throughout the sequence tracking.

*Optimization.* We optimize the objective function Equation (2) using Adam [42] in PyTorch. While recent soft-rasterizers [48, 55] are popular, we rely on a sampling based scheme as introduced by Thies et al. [70] to implement the differentiable rendering for the photo-metric reproduction error  $E_{dense}(\phi)$ . Specifically, we use a classical rasterizer to render the surface of the current estimation. The rasterized surface points that survive the depth test are considered as the set of visible surface points  $\mathcal{V}$  for which we compute the energy term  $E_{dense}(\phi) = \sum_{i \in \mathcal{V}} |I(\pi(\mathbf{R} \cdot p_i(\phi) + \mathbf{t})) - c_i(\phi)|$  where  $p_i$  and  $c_i$  being the  $i$ -th vertex and color of the reconstructed model, and  $I$  the RGB input image.

## 5 Dataset Unification

In the past, methods and their training scheme were limited by the availability of 3D scan datasets of human faces. While several small and medium-scale datasets are available, they are in different formats and do not share the same topology. To this end, we unified the available datasets such that they can be used as a supervision signal for face reconstruction from 2D images. Specifically, we register the FLAME [47] head model to the provided scan data. In an initial step, we fit the model to landmarks and optimize for the FLAME parameters based on an iterative closest point (ICP) scheme [5]. We further jointly optimize FLAME’s

Table 1: Overview of our unified datasets. The used datasets vary in the capture modality and the capture protocol. Here, we list the number of subject, the minimum number of images per subjects, and whether the dataset includes facial expressions. In total our dataset contains 2315 subjects with FLAME topology.

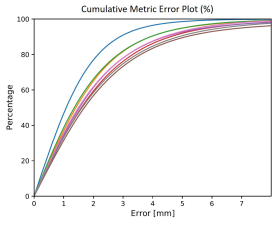
Dataset		#Subj.	#Min. Img.	Expr.
Stirling [26]		133	8	✓
D3DFACS [15]		10	videos	✓
Florence 2D/3D [3]		53	videos	✓
BU-3DFE [81]		100	83	✓
LYHM [16]		1211	2	✗
FaceWarehouse [10]		150	119	✓
FRGC [53]		531	7	✓
BP4D+ [82]		127	videos	✓

model parameters, and refine the fitting with a non-rigid deformation regularized by FLAME, similar to Li and Bolkart et al. [47]. In Table 1, we list the datasets that we unified for this project. We note that the datasets vary in the capturing modality and capturing script (with and without facial expressions, with and without hair caps, indoor and outdoor imagery, still images, and videos), which is suitable for generalization. The datasets are recorded in different regions of the world and are often biased towards ethnicity. Thus, combining other datasets results in a more diverse data pool. In the supplemental document, we show an ablation on the different datasets. *Upon agreement of the different dataset owners, we will share our unified dataset, i.e., for each subject one registered mesh with neutral expression in FLAME topology.* Note that in addition to the datasets listed in Table 1, we analyzed the FaceScape dataset [80]. While it provides a large set of 3D reconstructions ( $\sim 17k$ ), which would be ideal for our training, the reconstructions are not done in a metrical space. Specifically, the data has been captured in an uncalibrated setup and faces are normalized by the eye distance, which has not been detailed in their paper (instead, they mention sub-millimeter reconstruction accuracy which is not valid). This is a fundamental flaw of this dataset, and also questions their reconstruction benchmark [83].

## 6 Results

Our experiments mainly focus on the metrical reconstruction of a human face from in the wild images. In the supplemental document, we show results for the sequential tracking of facial motions using our metrical reconstruction as initialization. The following experiments are conducted with the original models of the respective publications including their reconstructions submitted to the given benchmarks. Note that these models are trained on their large-scale datasets, training them on our medium-scale 3D dataset would lead to overfitting.

Table 2: Quantitative evaluation of the face shape estimation on the *NoW Challenge* [59]. Note that we list two different evaluations: the non-metrical evaluation from the original NoW challenge and our new metrical evaluation (including a cumulative error plot on the left). The original NoW challenge cannot be considered metrical since Procrustes analysis is used to align the reconstructions to the corresponding reference meshes, including scaling. We list all methods from the original benchmark and additionally show the performance of the average human face of FLAME [47] as a reference (first row).

NoW-Metric Challenge		Non-Metrical [59]			Metrical (mm)		
	Method	Median	Mean	Std	Median	Mean	Std
	Average Face (FLAME [47])	1.21	1.53	1.31	1.49	1.92	1.68
	3DMM-CNN [72]	1.84	2.33	2.05	3.91	4.84	4.02
	PRNet [24]	1.50	1.98	1.88	—	—	—
	Deng et al [18] (TensorFlow)	1.23	1.54	1.29	2.26	2.90	2.51
	Deng et al [18] (PyTorch)	1.11	1.41	1.21	1.62	2.21	2.08
	RingNet [59]	1.21	1.53	1.31	1.50	1.98	1.77
	3DDFA-V2 [34]	1.23	1.57	1.39	1.53	2.06	1.95
	MGCNet [62]	1.31	1.87	2.63	1.70	2.47	3.02
	UMDFA [43]	1.52	1.89	1.57	2.31	2.97	2.57
	Dib et al. [19]	1.26	1.57	1.31	1.59	2.12	1.93
	DECA [23]	1.09	1.38	1.18	1.35	1.80	1.64
	FOCUS [46]	1.04	1.30	1.10	1.41	1.85	1.70
	• Ours	<b>0.90</b>	<b>1.11</b>	<b>0.92</b>	<b>1.08</b>	<b>1.37</b>	<b>1.17</b>

## 6.1 Face Shape Estimation

In recent publications, face shape estimation is evaluated on datasets where reference scans of the subjects are available. The NoW Challenge [59] and the benchmark of Feng et al. [25] which is based on Stirling meshes [26] are used in the state-of-the-art methods [18, 23, 59]. We conduct several studies on these benchmarks and propose different evaluation protocols.

**Non-Metrical Benchmark.** The established evaluation methods on these datasets are based on an optimal scaling step, i.e., to align the estimation to the reference scan, they optimize for a rigid alignment and an additional scaling factor which results in a non-metric/relative error. This scaling compensates for shape mispredictions, e.g., the mean error evaluated on the NoW Challenge for the average FLAME mesh (Table 2) drops from 1.92mm to 1.53mm because of the applied scale optimization. This is an improvement of around 20% which has nothing to do with the reconstruction quality and, thus, creates a misleading benchmark score where methods appear better than they are. Nevertheless, we evaluate our method on these benchmarks and significantly outperform all state-of-the-art methods as can be seen in Tables 2 and 4 (‘Non-Metrical’ column).

**Metrical Benchmark.** Since for a variety of applications, actual metrical reconstructions are required, we argue for a new evaluation scheme that uses a purely rigid alignment, i.e., without scale optimization (see Figure 5). The error is calculated using an Euclidean distance between each scan vertex and the closest point on the mesh surface. This new evaluation scheme enables a comparison of methods based on metrical quantities (see Tables 2 and 4) and, thus,



Table 3: Quantitative evaluation of the face shape estimation on the *Stirling Reconstruction Benchmark* [25] using the NoW protocol [59]. We list two different evaluations: the non-metric evaluation from the original benchmark and the metric evaluation. *Note that for this experiment, we exclude the Stirling dataset from our training set.*

Stirling (NoW Protocol)	Non-Metrical						Metrical (mm)					
	Median		Mean		Std		Median		Mean		Std	
	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ
Average Face (FLAME [47])	1.23	1.22	1.56	1.55	1.38	1.35	1.44	1.40	1.84	1.79	1.64	1.57
RingNet [59]	1.17	1.15	1.49	1.46	1.31	1.27	1.37	1.33	1.77	1.72	1.60	1.54
3DDFA-V2 [34]	1.26	1.20	1.63	1.55	1.52	1.45	1.49	1.38	1.93	1.80	1.78	1.68
Deng et al. [18] (TensorFlow)	1.22	1.13	1.57	1.43	1.40	1.25	1.85	1.81	2.41	2.29	2.16	1.97
Deng et al. [18] (PyTorch)	1.12	0.99	1.44	1.27	1.31	1.15	1.47	1.31	1.93	1.71	1.77	1.57
DECA [23]	1.09	1.03	1.39	1.32	1.26	1.18	1.32	1.22	1.71	1.58	1.54	1.42
<b>Ours w/o. Stirling</b>	<b>0.96</b>	<b>0.92</b>	<b>1.22</b>	<b>1.16</b>	<b>1.11</b>	<b>1.04</b>	<b>1.15</b>	<b>1.06</b>	<b>1.46</b>	<b>1.35</b>	<b>1.30</b>	<b>1.20</b>

Table 4: Quantitative evaluation of the face shape estimation on the *Stirling Reconstruction Benchmark* [25]. We list two different evaluations: the non-metric evaluation from the original benchmark and the metric evaluation. This benchmark is based on an alignment protocol that only relies on reference landmarks and, thus, is very noisy and dependent on the landmark reference selection (in our evaluation, we use the landmark correspondences provided by the FLAME [47] model). We use the image file list from [59] to compute the scores (i.e., excluding images where a face is not detectable). *Note that for this experiment, we exclude the Stirling dataset from our training set.*

Stirling/ESRC Benchmark	Non-Metrical [25]						Metrical (mm)					
	Median		Mean		Std		Median		Mean		Std	
	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ
Average Face (FLAME [47])	1.58	1.62	2.06	2.08	1.82	1.83	1.70	1.62	2.19	2.09	1.96	1.85
RingNet [59]	1.56	1.60	2.01	2.05	1.75	1.76	1.67	1.64	2.16	2.09	1.90	1.81
3DDFA-V2 [34]	1.58	1.49	2.03	1.90	1.74	1.63	1.70	1.56	2.16	1.98	1.88	1.70
Deng et al. [18] (TensorFlow)	1.56	1.41	2.02	1.84	1.77	1.63	2.13	2.14	2.71	2.65	2.33	2.12
Deng et al. [18] (PyTorch)	1.51	1.29	1.95	1.64	1.71	1.39	1.78	1.54	2.28	1.97	1.97	1.68
DECA [23]	1.40	1.32	1.81	1.72	1.59	1.50	1.56	1.45	2.03	1.87	1.81	1.64
<b>Ours w/o. Stirling</b>	<b>1.26</b>	<b>1.22</b>	<b>1.62</b>	<b>1.55</b>	<b>1.41</b>	<b>1.34</b>	<b>1.36</b>	<b>1.26</b>	<b>1.73</b>	<b>1.60</b>	<b>1.48</b>	<b>1.37</b>

is *fundamentally* different from the previous evaluation schemes. In addition, the benchmark of Feng et al. [25] is based on the alignment using sparse facial (hand-selected) landmarks. Our experiments showed that this scheme is highly dependent on the selection of these markers and results in inconsistent evaluation results. In our listed results, we use the marker correspondences that come with the FLAME model [47]. To get a more reliable evaluation scheme, we evaluate the benchmark of Feng et al. using the dense iterative closest point (ICP) technique from the NoW challenge, see Table 3. On all metrics, our proposed method significantly improves the reconstruction accuracy. Note that some methods are even performing worse than the mean face [47].



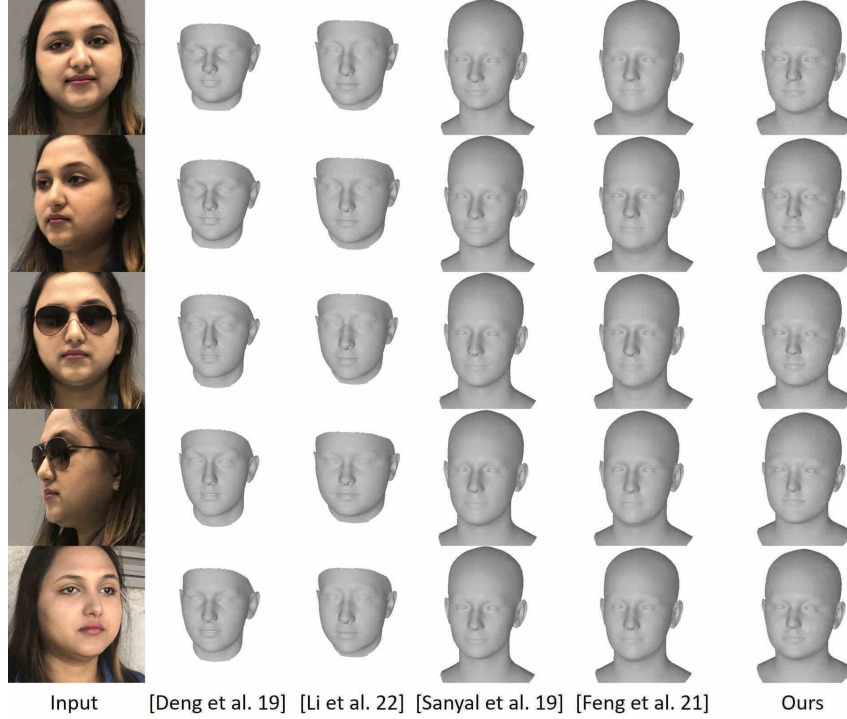


Fig. 3: Qualitative results on NoW Challenge [59] to show the invariance of our method to changes in illumination, expression, occlusion, rotation, and perspective distortion in comparison to other methods.

**Qualitative Results.** In Figure 3, we show qualitative results to analyze the stability of the face shape prediction of a subject across different expressions, head rotation, occlusions, or perspective distortion. As can be seen, our method is more persistent compared to others, especially, in comparison to Deng et al. [18] where shape predictions vary the most. Figure 4 depicts the challenging scenario of reconstructing toddlers from single images. Instead of predicting a small face for a child, the state of the art methods are predicting faces of adults. In contrast, MICA predicts the shape of a child with a correct scale.

In Figure 6 reconstructions for randomly sampled identities from the VoxCeleb2 [14] dataset are shown. Some of the baselines, especially, RingNet [59], exhibits strong bias towards the mean human face. In contrast, our method is able to not only predict better overall shape but also to reconstruct challenging regions like nose or chin, even though the training dataset contains a much smaller identity and ethnicity pool. Note that while the reconstructions of the baseline methods look good under the projection, they are not metric as shown in Tables 2 and 4.

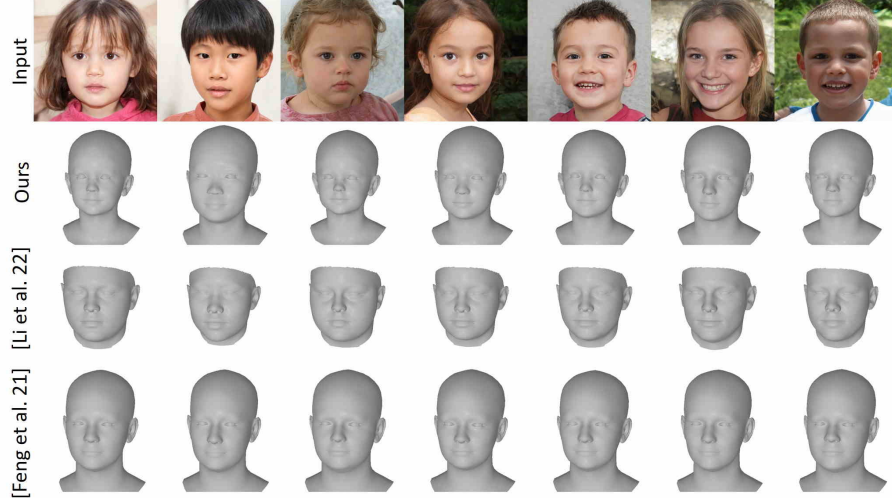


Fig. 4: Current methods are not predicting metrical faces, which becomes visible when displaying them in a metrical space and not in their image spaces. To illustrate we render the prediction of the faces of toddlers in a common metrical space using the same projection. State-of-the-art approaches trained in a self-supervised fashion like DECA [23] or weakly-supervised like FOCUS [46] scale the face of an adult to fit the observation in the image space, thus, the prediction in 3D is non-metrical. In contrast, our reconstruction method is able to recover the physiognomy of the toddlers. Input images are generated by StyleGan2 [38].

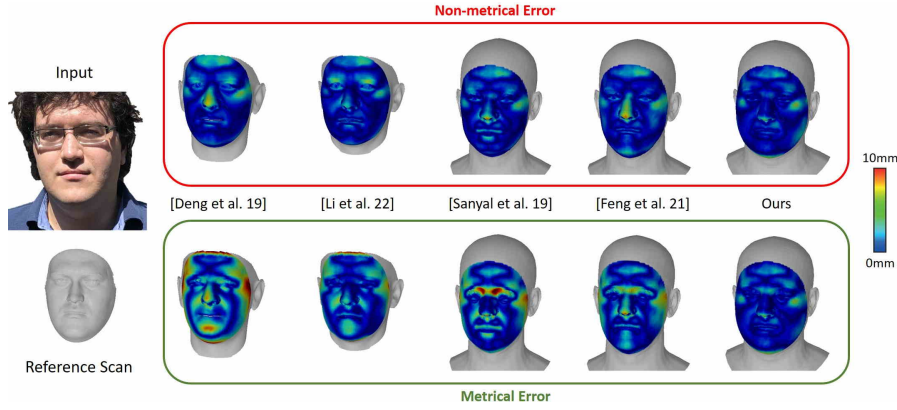


Fig. 5: Established evaluation benchmarks like [25, 59] are based on a non-metrical error metric (top-row). We propose a new evaluation protocol which measures reconstruction errors in a metrical space (bottom row) (c.f. Table 2). Image from the NoW [59] validation set.

## 6.2 Limitations

Our method is not designed to predict shape and expressions in one forward pass, instead, we reconstruct the expression separately using an optimization-based tracking method. However, this optimization-based tracking leads to temporally coherent results, as can be seen in the suppl. video. In contrast to DECA [23] or Deng et al. [18], the focus of our method is the reconstruction of a metrical 3D model, reconstructing high-frequent detail on top of our prediction is an interesting future direction. Our method fails, when the used face detector [61] does not recognize a face in the input.

## 7 Discussion & Conclusion

A metrical reconstruction is key for any application that requires the measurement of distances and dimensions. It is essential for the composition of reconstructed humans and scenes where objects of known size are in, thus, it is especially important for virtual reality and augmented reality applications. However, we show that recent methods and evaluation schemes are not designed for this task. While the established benchmarks report numbers in millimeters, they are computed with an optimal scale to align the prediction and the reference. We strongly argue against this practice, since it is misleading and the errors are not absolute metrical measurements. To this end, we propose a simple, yet fundamental adjustment of the benchmarks to enable metrical evaluations. Specifically, we remove the optimal scaling, and only allow rigid alignment of the prediction with the reference shape. As a stepping stone towards metrical reconstructions, we unified existing small- and medium-scale datasets of paired 2D/3D data. This allows us to establish 3D supervised losses in our novel shape prediction framework. While our data collection is still comparably small (around 2k identities), we designed MICA that uses features from a face recognition network pretrained on a large-scale 2D image dataset to generalize to in-the-wild image data. We validated our approach in several experiments and show state-of-the-art results on our newly introduced metrical benchmarks as well as on the established scale-invariant benchmarks. We hope that this work inspires researchers to concentrate on metrical face reconstruction.

*Acknowledgement.* We thank Haiwen Feng for support with NoW and Stirling evaluations, and Chunlu Li for providing FOCUS results. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Wojciech Zielonka.

*Disclosure.* While TB is part-time employee of Amazon, his research was performed solely at, and funded solely by MPI. JT is supported by Microsoft Research gift funds.

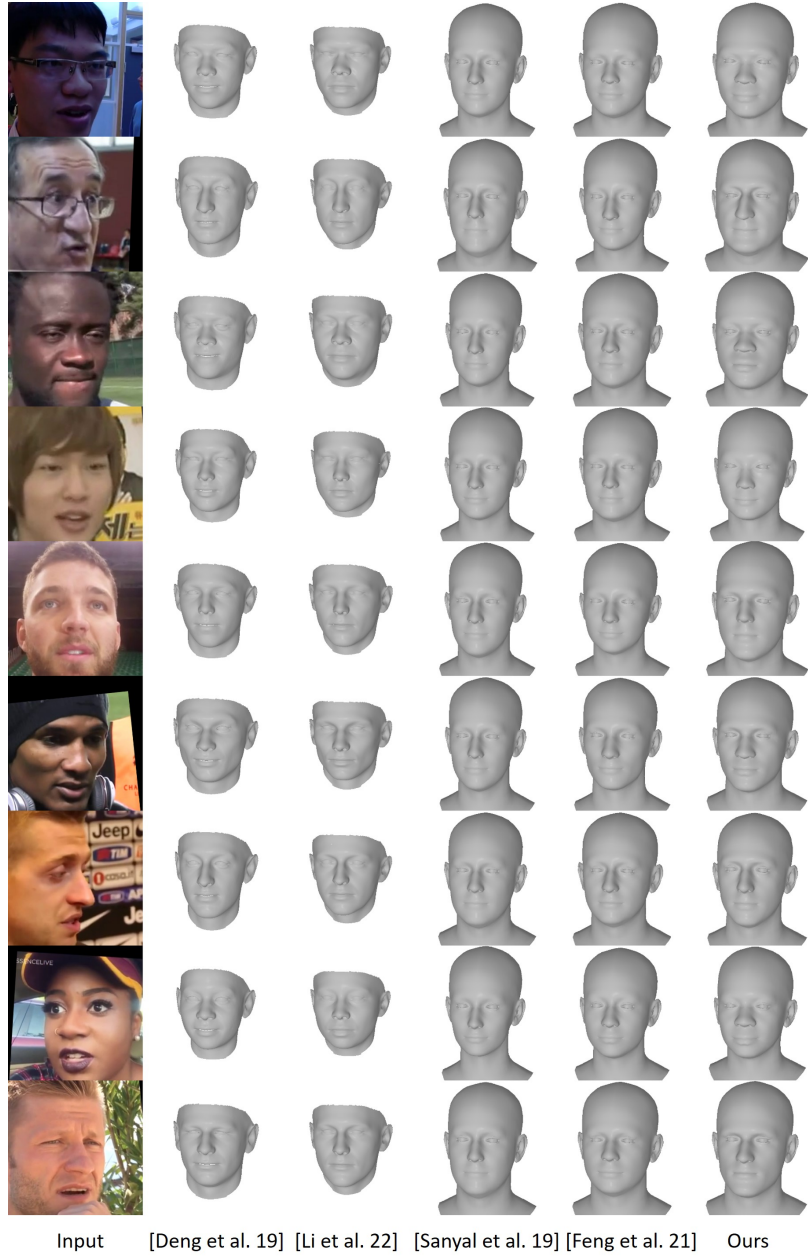


Fig. 6: Qualitative comparison on randomly sampled images from the VoxCeleb2 [14] dataset. Our method is able to capture face shape with intricate details like nose and chin, while being metrical plausible (c.f., Tables 2 and 4).

## Bibliography

- [1] Abrevaya, V.F., Boukhayma, A., Torr, P.H., Boyer, E.: Cross-modal deep face normals with deactivable skip connections. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4978–4988 (2020) [4](#)
- [2] An, X., Zhu, X., Xiao, Y., Wu, L., Zhang, M., Gao, Y., Qin, B., Zhang, D., Ying, F.: Partial fc: Training 10 million identities on a single machine. In: Arxiv 2010.05222 (2020) [5](#)
- [3] Bagdanov, A.D., Del Bimbo, A., Masi, I.: The florence 2D/3D hybrid face dataset. In: Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding. p. 79–80. J-HGBU '11, Association for Computing Machinery, New York, NY, USA (2011). <https://doi.org/10.1145/2072572.2072597>, <https://doi.org/10.1145/2072572.2072597> [6](#), [8](#)
- [4] Bas, A., Smith, W.A.P.: What does 2D geometric information really tell us about 3D face shape? International Journal of Computer Vision (IJCV) **127**(10), 1455–1473 (2019) [2](#)
- [5] Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures. vol. 1611, pp. 586–606. International Society for Optics and Photonics (1992) [7](#)
- [6] Blanz, V., Basso, C., Poggio, T., Vetter, T.: Reanimating faces in images and video. In: EUROGRAPHICS (EG). vol. 22, pp. 641–650 (2003) [3](#)
- [7] Blanz, V., Scherbaum, K., Vetter, T., Seidel, H.P.: Exchanging faces in images. Computer Graphics Forum **23**(3), 669–676 (2004) [3](#)
- [8] Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: SIGGRAPH. pp. 187–194 (1999) [3](#), [4](#)
- [9] Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: International Conference on Computer Vision (2017) [7](#)
- [10] Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: FaceWarehouse: A 3D facial expression database for visual computing. Transactions on Visualization and Computer Graphics **20**, 413–425 (01 2013) [2](#), [8](#)
- [11] Chang, F.J., Tran, A.T., Hassner, T., Masi, I., Nevatia, R., Medioni, G.: Expnet: Landmark-free, deep, 3d facial expressions. In: International Conference on Automatic Face & Gesture Recognition (FG). pp. 122–129 (2018) [4](#)
- [12] Chaudhuri, B., Vedapant, N., Shapiro, L., Wang, B.: Personalized face modeling for improved face reconstruction and motion retargeting (2020) [4](#)
- [13] Chen, A., Chen, Z., Zhang, G., Mitchell, K., Yu, J.: Photo-realistic facial details synthesis from single image. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9429–9439 (2019) [5](#)
- [14] Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: Deep speaker recognition. In: "INTERSPEECH" (2018) [11](#), [14](#)

- [15] Cosker, D., Krumhuber, E., Hilton, A.: A faces valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In: 2011 International Conference on Computer Vision. pp. 2296–2303 (2011). <https://doi.org/10.1109/ICCV.2011.6126510> 8
- [16] Dai, H., Pears, N., Smith, W., Duncan, C.: Statistical modeling of cranio-facial shape and texture. *International Journal of Computer Vision (IJCV)* **128**(2), 547–571 (2019). <https://doi.org/10.1007/s11263-019-01260-7> 2, 8
- [17] Deng, J., Guo, J., Liu, T., Gong, M., Zafeiriou, S.: Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In: European Conference on Computer Vision (ECCV). pp. 741–757 (2020) 3, 5
- [18] Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W) (2019) 2, 5, 9, 10, 11, 13
- [19] Dib, A., Thebault, C., Ahn, J., Gosselin, P., Theobalt, C., Chevallier, L.: Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In: International Conference on Computer Vision (ICCV). pp. 12819–12829 (2021) 9
- [20] Dou, P., Shah, S.K., Kakadiaris, I.A.: End-to-end 3d face reconstruction with deep neural networks (2017) 4
- [21] Egger, B., Smith, W.A.P., Tewari, A., Wuhler, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., Theobalt, C., Blanz, V., Vetter, T.: 3D morphable face models - past, present and future. *Transactions on Graphics (TOG)* **39**(5) (2020). <https://doi.org/10.1145/3395208> 3
- [22] Feng, H., Bolkart, T.: Photometric FLAME fitting (2020), [https://github.com/HavenFeng/photometric\\_optimization](https://github.com/HavenFeng/photometric_optimization) 7
- [23] Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. *Transactions on Graphics, (Proc. SIGGRAPH)* **40**(8) (2021) 2, 4, 5, 9, 10, 12, 13
- [24] Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3D face reconstruction and dense alignment with position map regression network. In: European Conference on Computer Vision (ECCV). pp. 534–551 (2018) 4, 9
- [25] Feng, Z., Huber, P., Kittler, J., Hancock, P.J.B., Wu, X., Zhao, Q., Koppen, P., Räscher, M.: Evaluation of dense 3D reconstruction from 2D face images in the wild. In: International Conference on Automatic Face & Gesture Recognition (FG). pp. 780–786 (2018). <https://doi.org/10.1109/FG.2018.00123> 2, 9, 10, 12
- [26] Feng, Z., Huber, P., Kittler, J., Hancock, P.J.B., Wu, X., Zhao, Q., Koppen, P., Räscher, M.: Evaluation of dense 3d reconstruction from 2d face images in the wild. *CoRR* **abs/1803.05536** (2018), <http://arxiv.org/abs/1803.05536> 2, 8, 9
- [27] Garrido, P., Valgaerts, L., Rehmsen, O., Thormaehlen, T., Perez, P., Theobalt, C.: Automatic face reenactment. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4217–4224 (2014) 3



- [28] Garrido, P., Valgaerts, L., Sarmadi, H., Steiner, I., Varanasi, K., Perez, P., Theobalt, C.: VDub - modifying face video of actors for plausible visual alignment to a dubbed audio track. In: EUROGRAPHICS (EG). pp. 193–204 (2015) [3](#)
- [29] Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) [5](#)
- [30] Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.P.: Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021) [5](#)
- [31] Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlastic, D., Freeman, W.T.: Unsupervised training for 3d morphable model regression (2018) [4](#)
- [32] Grassal, P.W., Prinzler, M., Leistner, T., Rother, C., Nießner, M., Thies, J.: Neural head avatars from monocular rgb videos (2021) [4](#)
- [33] Grishchenko, I., Ablavatski, A., Kartynnik, Y., Raveendran, K., Grundmann, M.: Attention mesh: High-fidelity face mesh prediction in real-time (2020) [7](#)
- [34] Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: European Conference on Computer Vision (ECCV) (2020) [9](#), [10](#)
- [35] Güler, R.A., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., Kokkinos, I.: Densereg: Fully convolutional dense shape regression in-the-wild (2017) [4](#)
- [36] Hu, L., Saito, S., Wei, L., Nagano, K., Seo, J., Fursund, J., Sadeghi, I., Sun, C., Chen, Y.C., Li, H.: Avatar digitization from a single image for real-time rendering. *ACM Trans. Graph.* **36**(6) (nov 2017). <https://doi.org/10.1145/3130800.31310887>, <https://doi.org/10.1145/3130800.31310887> [4](#)
- [37] Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G.: Large pose 3d face reconstruction from a single image via direct volumetric cnn regression (2017) [4](#)
- [38] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proc. CVPR (2020) [1](#), [12](#)
- [39] Kartynnik, Y., Ablavatski, A., Grishchenko, I., Grundmann, M.: Real-time facial surface geometry from monocular video on mobile gpus (2019) [7](#)
- [40] Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. *Transactions on Graphics (TOG)* **37**(4), 1–14 (2018) [3](#), [4](#)
- [41] Kim, H., Zollhöfer, M., Tewari, A., Thies, J., Richardt, C., Theobalt, C.: InverseFaceNet: Deep monocular inverse face rendering. In: Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2018) [4](#)
- [42] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *CoRR abs/1412.6980* (2015) [7](#)

- [43] Koizumi, T., Smith, W.A.P.: "look ma, no landmarks!" - unsupervised, model-based dense face alignment. In: European Conference on Computer Vision (ECCV). vol. 12347, pp. 690–706 (2020) [9](#)
- [44] Lattas, A., Moschoglou, S., Gecer, B., Ploumpis, S., Triantafyllou, V., Ghosh, A., Zafeiriou, S.: AvatarMe: Realistically renderable 3D facial reconstruction "in-the-wild". In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 760–769 (2020) [5](#)
- [45] Lattas, A., Moschoglou, S., Ploumpis, S., Gecer, B., Ghosh, A., Zafeiriou, S.P.: AvatarMe++: Facial shape and BRDF inference with photorealistic rendering-aware GANs. Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2021) [5](#)
- [46] Li, C., Morel-Forster, A., Vetter, T., Egger, B., Kortylewski, A.: To fit or not to fit: Model-based face reconstruction and occlusion segmentation from weak supervision. CoRR [abs/2106.09614](#) (2021), <https://arxiv.org/abs/2106.09614> [9](#), [12](#)
- [47] Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6), 194:1–194:17 (2017), <https://doi.org/10.1145/3130800.3130813> [2](#), [3](#), [6](#), [7](#), [8](#), [9](#), [10](#)
- [48] Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. International Conference on Computer Vision (ICCV) (Oct 2019) [7](#)
- [49] Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. CoRR [abs/1711.05101](#) (2017), <http://arxiv.org/abs/1711.05101> [6](#)
- [50] Morales, A., Piella, G., Sukno, F.M.: Survey on 3d face reconstruction from uncalibrated images (2021) [3](#)
- [51] Nagano, K., Seo, J., Xing, J., Wei, L., Li, Z., Saito, S., Agarwal, A., Fursund, J., Li, H.: Pagan: Real-time avatars using dynamic textures. ACM Trans. Graph. **37**(6) (dec 2018). <https://doi.org/10.1145/3272127.3275075>, <https://doi.org/10.1145/3272127.3275075> [4](#)
- [52] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3D face model for pose and illumination invariant face recognition. In: International conference on advanced video and signal based surveillance. pp. 296–301 (2009) [3](#)
- [53] Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 947–954 vol. 1 (2005). <https://doi.org/10.1109/CVPR.2005.268> [8](#)
- [54] Ramamoorthi, R., Hanrahan, P.: An efficient representation for irradiance environment maps. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. p. 497–500. SIGGRAPH '01, Association for Computing Machinery, New York, NY, USA (2001). <https://doi.org/10.1145/383259.383317>, <https://doi.org/10.1145/383259.383317> [7](#)



- [55] Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d. arXiv:2007.08501 (2020) 7
- [56] Richardson, E., Sela, M., Kimmel, R.: 3d face reconstruction by learning from synthetic data (2016) 4
- [57] Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image (2017) 4
- [58] Saito, S., Wei, L., Hu, L., Nagano, K., Li, H.: Photorealistic facial texture inference using deep neural networks (2016) 5
- [59] Sanyal, S., Bolkart, T., Feng, H., Black, M.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1, 2, 4, 5, 9, 10, 11, 12
- [60] Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2015). <https://doi.org/10.1109/cvpr.2015.7298682>, <http://dx.doi.org/10.1109/CVPR.2015.7298682> 4, 5
- [61] Serengil, S.I., Ozpinar, A.: Hyperextended lightface: A facial attribute analysis framework. In: 2021 International Conference on Engineering and Emerging Technologies (ICEET). pp. 1–4. IEEE (2021). <https://doi.org/10.1109/ICEET53442.2021.9659697>, <https://doi.org/10.1109/ICEET53442.2021.9659697> 13
- [62] Shang, J., Shen, T., Li, S., Zhou, L., Zhen, M., Fang, T., Quan, L.: Self-supervised monocular 3D face reconstruction by occlusion-aware multi-view geometry consistency. In: European Conference on Computer Vision (ECCV). vol. 12360, pp. 53–70 (2020) 9
- [63] Tewari, A., Bernard, F., Garrido, P., Bharaj, G., Elgharib, M., Seidel, H.P., Pérez, P., Zollhofer, M., Theobalt, C.: Fml: Face model learning from videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10812–10822 (2019) 4
- [64] Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., Theobalt, C.: Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4
- [65] Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Christian, T.: MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In: The IEEE International Conference on Computer Vision (ICCV) (2017) 4
- [66] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Facevr: Real-time gaze-aware facial reenactment in virtual reality. ACM Transactions on Graphics 2018 (TOG) (2018) 3
- [67] Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: Audio-driven facial reenactment. European Conference on Computer Vision (ECCV) (2020) 4

- [68] Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. *Transactions on Graphics (TOG)* **38**(4), 1–12 (2019) [4](#)
- [69] Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. *Transactions on Graphics (TOG)* **34**(6) (2015) [4](#)
- [70] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2Face: Real-time face capture and reenactment of RGB videos. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2387–2395 (2016) [3](#), [4](#), [7](#)
- [71] Thies, J., Zollhöfer, M., Theobalt, C., Stamminger, M., Niessner, M.: Headon: Real-time reenactment of human portrait videos. *ACM Transactions on Graphics* **37**(4), 1–13 (Aug 2018). <https://doi.org/10.1145/3197517.3201350>, <http://dx.doi.org/10.1145/3197517.3201350> [3](#), [4](#)
- [72] Tran, A.T., Hassner, T., Masi, I., Medioni, G.: Regressing robust and discriminative 3D morphable models with a very deep neural network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1599–1608 (2017) [4](#), [9](#)
- [73] Tran, A.T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., Medioni, G.: Extreme 3D face reconstruction: Seeing through occlusions. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2018) [5](#)
- [74] Tran, L., Liu, F., Liu, X.: Towards high-fidelity nonlinear 3d face morphable model. In: *In Proceeding of IEEE Computer Vision and Pattern Recognition*. Long Beach, CA (June 2019) [4](#)
- [75] Tu, X., Zhao, J., Jiang, Z., Luo, Y., Xie, M., Zhao, Y., He, L., Ma, Z., Feng, J.: Joint 3D face reconstruction and dense face alignment from a single image with 2D-assisted self-supervised learning. *arXiv preprint arXiv:1903.09359* (2019) [4](#)
- [76] Wei, H., Liang, S., Wei, Y.: 3d dense face alignment via graph convolution networks (2019) [4](#)
- [77] Weise, T., Bouaziz, S., Li, H., Pauly, M.: Realtime performance-based facial animation. In: *Transactions on Graphics (TOG)*. vol. 30 (2011) [3](#)
- [78] Weise, T., Li, H., Gool, L.J.V., Pauly, M.: Face/Off: live facial puppetry. In: *SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*. pp. 7–16 (2009) [3](#)
- [79] Yamaguchi, S., Saito, S., Nagano, K., Zhao, Y., Chen, W., Olszewski, K., Morishima, S., Li, H.: High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Trans. Graph.* **37**(4) (jul 2018). <https://doi.org/10.1145/3197517.3201364>, <https://doi.org/10.1145/3197517.3201364> [4](#), [5](#)
- [80] Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020) [8](#)

- [81] Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.: A 3d facial expression database for facial behavior research. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR06). pp. 211–216 (2006). <https://doi.org/10.1109/FGR.2006.6> 8
- [82] Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., Cohn, J.F., Ji, Q., Yin, L.: Multimodal spontaneous emotion corpus for human behavior analysis. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3438–3446 (2016). <https://doi.org/10.1109/CVPR.2016.374> 8
- [83] Zhu, H., Yang, H., Guo, L., Zhang, Y., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. arXiv preprint arXiv:2111.01082 (2021) 2, 8
- [84] Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3D solution. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 146–155. IEEE Computer Society, Los Alamitos, CA, USA (jun 2016). <https://doi.org/10.1109/CVPR.2016.23>, <https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.23> 4
- [85] Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., Theobalt, C.: State of the art on monocular 3D face reconstruction, tracking, and applications. Computer Graphics Forum (Eurographics State of the Art Reports) **37**(2) (2018) 3

## A.2 INSTANT VOLUMETRIC HEAD AVATARS

*Instant Volumetric Head Avatars*

Wojciech Zielonka, Timo Bolkart, Justus Thies

Published in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, 2023.

**Abstract**

We present Instant Volumetric Head Avatars (INSTA), a novel approach for reconstructing photo-realistic digital avatars instantaneously. INSTA models a dynamic neural radiance field based on neural graphics primitives embedded around a parametric face model. Our pipeline is trained on a single monocular RGB portrait video that observes the subject under different expressions and views. While state-of-the-art methods take up to several days to train an avatar, our method can reconstruct a digital avatar in less than 10 minutes on modern GPU hardware, which is orders of magnitude faster than previous solutions. In addition, it allows for the interactive rendering of novel poses and expressions. By leveraging the geometry prior of the underlying parametric face model, we demonstrate that INSTA extrapolates to unseen poses. In quantitative and qualitative studies on various subjects, INSTA outperforms state-of-the-art methods regarding rendering quality and training time.

# Instant Volumetric Head Avatars

Wojciech Zielonka    Timo Bolkart    Justus Thies

Max Planck Institute for Intelligent Systems, Tübingen, Germany

{wojciech.zielonka, timo.bolkart, justus.thies}@tuebingen.mpg.de

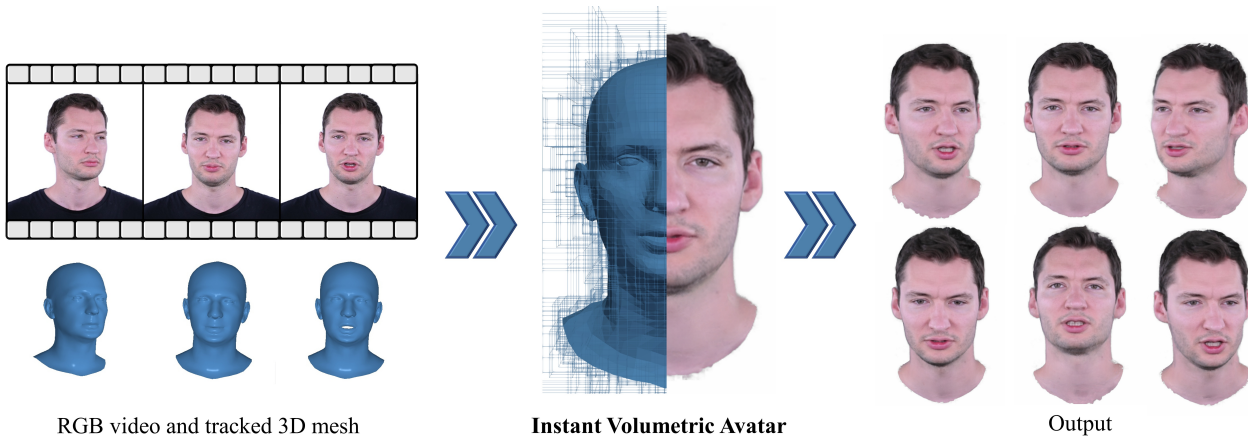


Figure 1. Given a short monocular RGB video, our method instantaneously optimizes a deformable neural radiance field to synthesize a photo-realistic animatable 3D neural head avatar. The neural radiance field is embedded in a multi-resolution grid around a 3D face model which guides the deformations. The resulting head avatar can be viewed under novel views and animated at interactive frame rates.

## Abstract

We present *Instant Volumetric Head Avatars (INSTA)*, a novel approach for reconstructing photo-realistic digital avatars instantaneously. INSTA models a dynamic neural radiance field based on neural graphics primitives embedded around a parametric face model. Our pipeline is trained on a single monocular RGB portrait video that observes the subject under different expressions and views. While state-of-the-art methods take up to several days to train an avatar, our method can reconstruct a digital avatar in less than 10 minutes on modern GPU hardware, which is orders of magnitude faster than previous solutions. In addition, it allows for the interactive rendering of novel poses and expressions. By leveraging the geometry prior of the underlying parametric face model, we demonstrate that INSTA extrapolates to unseen poses. In quantitative and qualitative studies on various subjects, INSTA outperforms state-of-the-art methods regarding rendering quality and training time. Project website: <https://zielon.github.io/insta/>

## 1. Introduction

For immersive telepresence in AR or VR, we aim for digital humans (avatars) that mimic the motions and facial expressions of the actual subjects participating in a meeting. Besides the motion, these avatars should reflect the human’s shape and appearance. Instead of prerecorded, old avatars, we aim to instantaneously reconstruct the subject’s look to capture the actual appearance during a meeting. To this end, we propose Instant Volumetric Head Avatars (INSTA), which enables the reconstruction of an avatar within a few minutes ( $\sim 10$  min) and can be driven at interactive frame rates. For easy accessibility, we rely on commodity hardware to train and capture the avatar. Specifically, we use a single RGB camera to record the input video. State-of-the-art methods that use similar input data to reconstruct a human avatar require a relatively long time to train, ranging from around one day [20] to almost a week [16, 58]. Our approach uses dynamic neural radiance fields [16] based on neural graphics primitives [38], which are embedded around a parametric face model [25], allowing low training times and fast evaluation. In contrast to existing methods, we use a metrical face reconstruction [59] to ensure that the avatar

has metrical dimensions such that it can be viewed in an AR/VR scenario where objects of known size are present. We employ a canonical space where the dynamic neural radiance field is constructed. Leveraging the motion estimation employing the parametric face model FLAME [25], we establish a deformation field around the surface using a bounding volume hierarchy (BVH) [12]. Using this deformation field, we map points from the deformed space into the canonical space, where we evaluate the neural radiance field. As the surface deformation of the FLAME model does not include details like wrinkles or the mouth interior, we condition the neural radiance field by the facial expression parameters. To improve the extrapolation to novel views, we further leverage the FLAME-based face reconstruction to provide a geometric prior in terms of rendered depth maps during training of the NeRF [36]. In comparison to state-of-the-art methods like NeRFace [16], IMAvatar [58], or Neural Head Avatars (NHA) [20], our method achieves a higher rendering quality while being significantly faster to train and evaluate. We quantify this improvement in a series of experiments, including an ablation study on our method.

In summary, we present Instant Volumetric Head Avatars with the following contributions:

- a surface-embedded dynamic neural radiance field based on neural graphics primitives, which allows us to reconstruct metrical avatars in a few minutes instead of hours or days,
- and a 3DMM-driven geometry regularization of the dynamic density field to improve pose extrapolation, an important aspect of AR/VR applications.

## 2. Related Work

INSTA is reconstructing animatable digital human avatars from monocular video data based on 3D neural rendering [48]. Current solutions are using implicit representations [8, 16, 29, 36, 40, 41] optimized via differentiable volumetric rendering, or are based on explicit models [5, 7, 20, 49] for instance, triangle or tetrahedral meshes using differentiable rasterization [10, 22, 30, 33]. For a concise overview of neural rendering methods and face reconstruction, we point the reader to the state-of-the-art reports by Zollhöfer et al. [60], and Tewari et al. [47, 48].

**Static Neural Radiance Fields.** Mildenhall et al. [36] and its many follow-up works [3, 4, 28, 35, 39, 44, 45, 51, 56], synthesize novel views of a complex static scene using differentiable volumetric rendering. Many methods suffer from a long training time (1-5 days). To this end, different acceleration methods have been proposed to improve the training time. Yu et al. [15] achieved  $100\times$  speedup by using a sparse voxel grid storing density and spherical harmonics coefficients at each node. The final color is the compo-

sition of tri-linearly interpolated values of each voxel intersecting with the ray. TensorRF [9] factorizes the 4D NeRF scene into multiple compact low-rank tensor components achieving high performance and compactness. The coordinate-based MLP is replaced with a voxel grid of features, and the final color is its vector-matrix outer product. Müller et al. [38] introduced a new computer graphics primitive in the form of tiny MLPs which benefit from a multi-resolution hashing encoding. The key idea is similar to Yu et al. [15]. The space is divided into an independent multi-level grid with feature vectors at the vertices of the grid. A spatial hash function [46] is used to store the voxel grid efficiently. Each point sampled on the ray is encoded by the interpolated feature vector of the corresponding grid level and passed to a tiny neural network to synthesize the final color. Our method uses this efficient architecture to model the face in a canonical space.

Some of the static NeRF methods [2, 13, 44, 52] use additional depth maps to improve alignment and quality for static scenes. The depth priors help guide the ray sampling and better estimate the transmittance, resulting in improved geometry and color recovery. While we are working with RGB images only, our method leverages the geometry prior of the 3DMM to guide the depth estimation during training, which results in an improved extrapolation ability w.r.t. view changes.

**Deformable Neural Radiance Fields.** After the introduction of NeRF [36] for static scenes, a natural research direction was to generalize it to dynamic, time-varying ones [14, 26, 40, 41, 43, 50]. The reconstruction problem is divided into two different spaces, the deformed scene, and the canonical space, with a neural network as the mapper between them. For human body modeling, a series of approaches have been proposed that leverage the kinematic chain of the SMPL [32] body model to condition the mapping function. Peng et al. [42] proposed to learn blend weights to estimate the linear blend skinning-based warping field between canonical and deformed space based on the body skeleton. Similarly, Neural Actor [29] uses a 3D body mesh proxy to learn pose-dependent geometric deformation and view-dependent appearance effects defined in the canonical space. Lombardi et al. [31], which defines surface-aligned neural volumes to improve the rendering speed. Garbin et al. [18] build a tetrahedral deformation graph around a radiance field based on the underlying mesh on which the deformations are defined, effectively transforming sampled points according to the current cage state. Xu et al. [53] propose surface-aligned neural radiance fields by projecting points in space to the surface of the body mesh. Our idea is based on a similar principle. However, instead of projecting points onto the mesh surface, we construct a 3D space around the head and deform points based on the deformation defined by the nearest triangles.



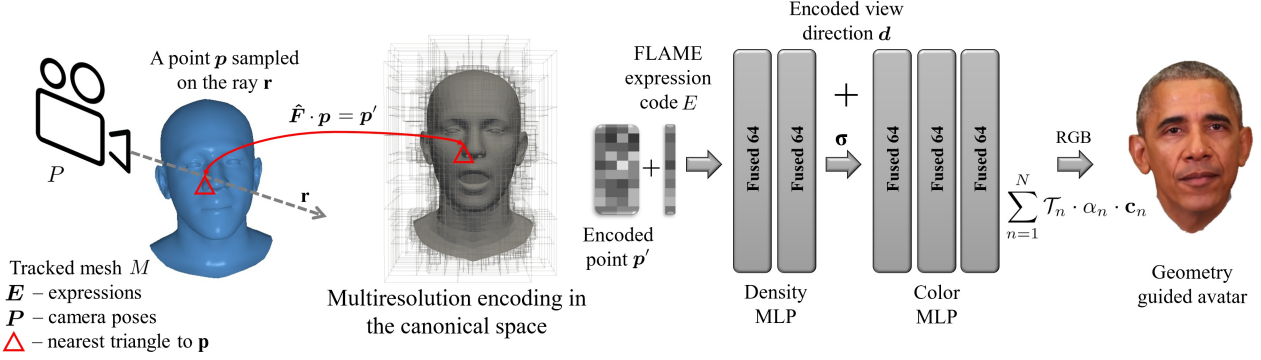


Figure 2. **Overview.** INSTA follows differentiable volumetric optimization introduced in [36, 38]. For each sampled point  $p \in \mathbb{R}^4$  in deformed space (in homogeneous coordinates), we are computing the nearest neighbor triangle on the mesh  $T_{def} \in M_i$  and its topological corresponding twin in the canonical space  $T_{canon} \in M^{canon}$ . The deformation gradient of the triangle from deformed space to canonical space  $\hat{F} \in \mathbb{R}^{4 \times 4}$  defines the deformation field. Specifically,  $p$  is transformed to the canonical space by  $p' = \hat{F} \cdot p$ . After canonicalization, the point is encoded using a multi-resolution hashing [38]. This feature is passed to fully fused multi-layer perceptrons [37] with additional conditioning on the facial expressions  $E_i$  and the encoded view direction  $d$ .

In contrast to modeling the deformation explicitly, Gafni et al. [16] implicitly model the facial expressions by conditioning the NeRF MLP with the global expression code obtained from 3DMM tracking [49] and by optimizing per latent frame codes to increase the network capacity for overfitting. In our approach, we leverage the idea of dynamic neural radiance fields to improve the mouth region’s rendering, which is not represented by the face model motion prior. Inspired by 3DMMs, IMAvatar [58] learns the subject-specific implicit representation of texture together with expression blendshapes and blend skinning weights. They optimize an implicit surface by incorporating ray marching from Yariv et al. [54] with root-finding of the occupancy function [11] to locate canonical correspondence of deformed points. However, we found the training time-consuming ( $\sim 5$  days) and unstable (can diverge). In a concurrent work, Gao et al. [17] create personalized blendshapes using neural graphics primitives, where for each of the blendshapes, a multi-resolution grid [38] is trained.

### 3. Instant Deformable Neural Radiance Field

Our goal is to create instant digital avatars which can be learned in a few minutes and rendered in interactive time. For this purpose, we are using a geometry-guided deformable neural radiance field embedded into a multi-resolution hashing grid [38], exploiting differentiable volumetric rendering [36] (see Fig. 2).

For a given monocular video consisting of images  $I = \{I_i\}$  along with optimized intrinsic camera parameters  $K \in \mathbb{R}^{3 \times 3}$ , tracked FLAME [25] meshes  $M = \{M_i\}$  with corresponding facial expression coefficients  $E = \{E_i\}$  and poses  $P = \{P_i\}$ , our goal is to build a controllable head

avatar represented by a neural radiance field. To this end, we employ a canonical space where the neural radiance field is constructed. To render specific facial expressions using volumetric rendering, we canonicalize the samples on a ray from the deformed space to query the radiance field in the canonical space.

**Volumetric Rendering.** We take advantage of the recent advances in interactive NeRF optimization and use neural graphic primitives [38] to represent the radiance field. The representation of the avatar is optimized using the differentiable volumetric rendering equation:

$$\hat{C} = \int_0^D \mathcal{T}(t) \cdot \sigma(t) \cdot c(t) dt + \mathcal{T}(D) \cdot c_{bg}, \quad (1)$$

where  $\mathcal{T}(t_n) = \exp\left(-\int_0^{t_n} \sigma(t) dt\right)$  is the transmittance which indicates the probability of a ray traveling from  $[0, t_n)$  without interaction with any other particles [36],  $\sigma(t)$  is the density and  $c(t)$  is the radiance at position  $p_t$ . Note that the sample points  $p_t$  are canonicalized to access the actual radiance field. Following NeRFace [16], we condition every sample  $p_t$  on the ray with the 3DMM facial expression code  $E_i \in \mathbb{R}^{16}$  of video frame  $i$ . Please note that in contrast to NeRFace [16] and IMAvatar [58], we do not use additional per-frame learnable codes. The viewing vector  $v \in \mathbb{R}^3$  is encoded using spherical harmonics projection on four basis functions [1, 38] resulting in the final viewing vector encoding  $d \in \mathbb{R}^{16}$  which is concatenated with density MLP output. While the viewing conditioning is applied on the entire avatar, the conditioning on facial expressions is bounded to the dynamically changing mouth region and is set to a constant vector  $E_i = \mathbf{1}$  for the other regions.

**Canonicalization.** We define a mapping function  $\Phi(\mathbf{p}, M_i)$  that projects a point  $\mathbf{p} \in \mathbb{R}^4$  from the time-varying deformed space (where the volumetric rendering is performed) to the canonical space. The mapping function leverages the time-varying surface approximation  $M_i$  and a predefined mesh in canonical space  $M^{canon}$ . We employ a nearest triangle search in deformed space to compute the deformation gradient  $\mathbf{F} \in \mathbb{R}^{4 \times 4}$  which is used to map point  $\mathbf{p}$  to the canonical counterpart  $\mathbf{p}'$ . The deformation gradient  $\mathbf{F}$  is computed via the known Frenet frames of the deformed triangle  $T_{def} \in M_i$  and the canonical triangle  $T_{canon} \in M^{canon}$ . Specifically, we compute the rotation matrices  $\{\mathbf{R}_{canon}, \mathbf{R}_{def}\} \in \mathbb{R}^{3 \times 3}$  based on the corresponding tangent, bitangent, and normal vectors of a triangle. With the translations  $\{\mathbf{t}_{canon}, \mathbf{t}_{def}\} \in \mathbb{R}^3$  defined by a vertex of the triangle, they form the Frenet coordinate system frames  $\mathbf{L}_{canon}$  and  $\mathbf{L}_{def} \in \mathbb{R}^{4 \times 4}$ :

$$\begin{aligned} \mathbf{L}_{def} &= \begin{bmatrix} \mathbf{R}_{def} & \mathbf{t}_{def} \\ \mathbf{0}^T & 1 \end{bmatrix}, \\ \mathbf{L}_{canon} &= \begin{bmatrix} \mathbf{R}_{canon} & \mathbf{t}_{canon} \\ \mathbf{0}^T & 1 \end{bmatrix}. \end{aligned} \quad (2)$$

To account for any potential triangle size change between deformed and canonical spaces, we compute an isotropic scaling factor  $\lambda \in \mathbb{R}$  via the relative surface area change of the given triangle w.r.t. its canonical twin  $\lambda = \frac{a_{def}}{a_{canon}}$ . The deformation gradient  $\mathbf{F}$  is defined as:

$$\begin{aligned} \mathbf{F} &= \mathbf{L}_{canon} \cdot \Lambda \cdot \mathbf{L}_{def}^{-1}, \\ \Lambda &= \begin{bmatrix} \lambda \mathbf{I} & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix}. \end{aligned} \quad (3)$$

To avoid transformation discontinuity, which arises from the local coordinate system of each triangle, we additionally perform exponentially weighted averaging of the transformations of the adjacent faces of the triangle's edges:

$$\hat{\mathbf{F}} = \frac{1}{\sum_{f \in A} \omega_f} \cdot \sum_{f \in A} \omega_f \mathbf{F}_f, \quad (4)$$

where  $\omega_f = \exp(-\beta \|\mathbf{c}_f - \mathbf{p}\|_2)$ ,  $\beta = 4$  and  $A$  is the set of adjacent faces to  $T$  (including  $T$  with  $\beta = 1$ ) with corresponding centroids  $\mathbf{c}_f$ . Please note that all vertex positions are defined in meters (FLAME metrical space).

To achieve interactive rendering as well as instantaneous optimization of the neural radiance field, we leverage a classical bounding volume hierarchy (BVH) [12] to significantly increase the nearest triangle search speed for the sampled points  $\mathbf{p}_t$  on the ray. Note that methods like IMAvatar [58] perform computation-heavy root-finding procedures to calculate surface points iteratively [11]. Our method builds a BVH based on the corresponding deformed mesh  $M_i$  of frame  $i$  to establish the mapping function to

the canonical mesh. Our BVH is implemented on GPU to utilize massively parallel nearest triangle search [23]. To alleviate the triangle search for highly tessellated FLAME regions, we simplified the eyeballs and the eye region [19]. Moreover, an additional set of triangles in the mouth region is used to serve as a deformation proxy (see sup. mat.).

### 3.1. Training Objectives

The optimization of the neural radiance field is based on a color reproduction objective and a geometry prior based on the 3DMM. Following NeRF [36], we redefine the volumetric rendering Equation (1) with piece-wise constant density and color, and rewrite it in terms of alpha-compositing:

$$\hat{\mathbf{C}}(t_{N+1}) = \sum_{n=1}^N \mathcal{T}_n \cdot \alpha_n \cdot \mathbf{c}_n, \quad (5)$$

where  $\mathcal{T}_n = \prod_{n=1}^{N-1} (1 - \alpha_n)$  weight  $\alpha_n$  is defined as  $\alpha_n \equiv 1 - \exp(-\sigma_n \delta_n)$  and  $\delta_n$  is a step size equal  $\frac{\sqrt{3}}{1024}$ . To measure the photometric error, we use a Huber loss [21] with  $\rho = 0.1$ :

$$\mathcal{L}_{color} = \begin{cases} \frac{1}{2}(\mathbf{C} - \hat{\mathbf{C}})^2 & \text{if } \|\mathbf{C} - \hat{\mathbf{C}}\| < \rho \\ \rho(\|\mathbf{C} - \hat{\mathbf{C}}\| - \frac{1}{2}\rho) & \text{otherwise} \end{cases} \quad (6)$$

We enforce a depth loss to leverage the geometry prior of the reconstructed face based on the 3DMM FLAME. Specifically, we rasterize the depth of the tracking mesh  $M_i$  and apply an L1 distance between this map and the ray termination of the volumetric rendering. As the FLAME model does not contain details like hair, we restrict the geometry prior to the face region:

$$\mathcal{L}_{geom} = \sum_{\mathbf{r}} |\mathbb{1}_{face}\{z(\mathbf{r}) - \hat{z}(\mathbf{r})\}|, \quad (7)$$

where  $\hat{z} = \sum_{n=1}^N \mathcal{T}_n \cdot \alpha_n \cdot t_n$ , and  $t_n$  is the current sample position, and  $\mathbb{1}_{face}\{\}$  is a segmentation indicator function which enables the loss for the face region. The  $\mathbb{1}_{face}$  function uses face parsing information [55] to decide a given pixel membership. The total loss  $\mathcal{L}$  is defined as:

$$\mathcal{L} = \sum_{\mathbf{r}} \lambda_{color}(\mathbf{r}) \mathcal{L}_{color}(\mathbf{r}) + \lambda_{geom} \mathcal{L}_{geom}(\mathbf{r}), \quad (8)$$

where  $\lambda_{geom} = 1.25$  controls the influence of the geometry prior and  $\lambda_{color}(\mathbf{r})$  weights the color loss contribution based on a face parsing mask. Specifically, we weight the color loss higher for the mouth region with  $\lambda_{color} = 40$  and  $\lambda_{color} = 1$  otherwise.

We implemented our animatable dynamic radiance field using the Nvidia NGP C++ framework [37]. We use two fully fused MLPs [37], each with 64 neurons, for color and



density predictions. The density MLP outputs feature values vector  $\sigma \in \mathbb{R}^{16}$  where the first value is the log-space density. The vector  $\sigma$  is later concatenated with the encoded viewing vector  $d$  to be the input of the color network. For optimization, we used Adam [24] with an exponential moving average on the weights and fixed learning rate  $\eta = 2.5e-3$ . In our experiments, we train the network for 32k optimization steps. We randomly sample 1700 frames from the whole dataset during the training and load them into the processing buffer. Every 1500 steps, we repeat the procedure and resample the dataset.

## 4. Dataset

Our method takes a single video as input to generate the volumetric avatar of the depicted subject. For our experiments, we recorded multiple actors with a Nikon Z6 II Camera as well as used sequences from Youtube, resulting in a set of twelve actors. For the in-house recordings, we captured around 2-3min of monocular RGB Full HD videos, which later were cropped, sub-sampled to 25fps, and resized to  $512^2$  resolution. We additionally use background foreground segmentation using robust matting [27] and an off-the-shelf face parsing framework [55] for image segmentation and clothes removal.

**Dataset Tracking Generation.** An essential part of this project is temporally stable face tracking of the monocular input data. To this end, we use the analysis-by-synthesis-based face tracker from MICA [59], based on Face2Face [49] using a sampling-based differentiable rendering. We refer to the original paper [49] for more details. We extend the optimization with two extra blendshapes for eyelids and iris tracking using Mediapipe [34]. In contrast to MICA, we also optimize for FLAME shape parameters, with regularization towards MICA shape prediction instead of the average face shape as in Face2Face [49]. Note that for our prototype, we implemented the tracking in PyTorch, which is significantly slower than the original Face2Face implementation, which can track faces in real-time.

## 5. Results

In this section, we evaluate the quality of the synthesized digital human avatars generated by our method INSTA in comparison to state-of-the-art. For this purpose, we use the test sequences from our dataset, which consist of the last 350 frames of each video.

### 5.1. Image Quality Evaluation

To evaluate our method in terms of the image quality and novel view extrapolation, we make a comparison to NeRFace [16], IMAvatar [58], and Neural Head Avatars (NHA) [20]. For this comparison, we use the original im-

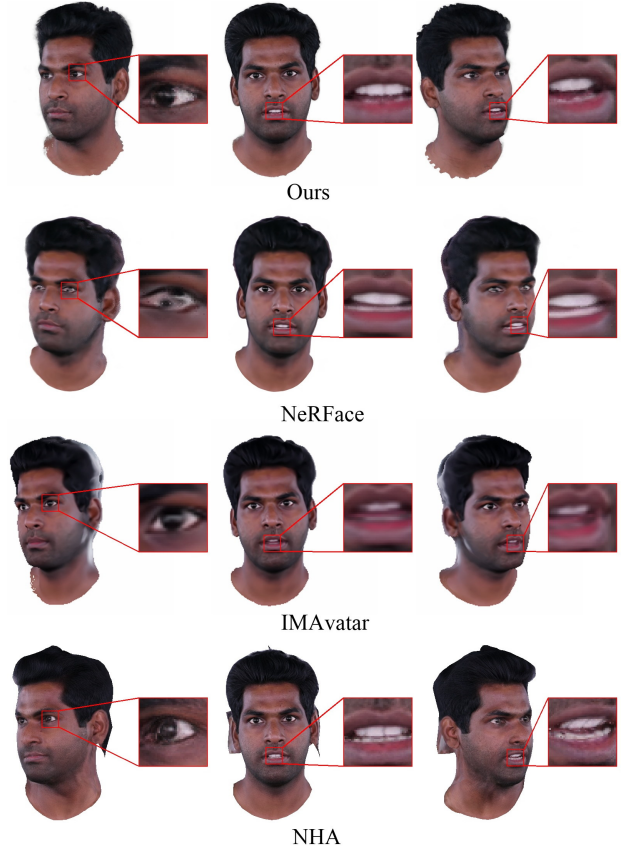


Figure 3. Qualitative comparison for novel view extrapolation. As can be seen, our method can better handle image synthesis under novel poses. NHA [20] suffers from degenerated geometry with many artifacts at the ear region. NeRFace [16] lacks high-frequency details for eyes and teeth, and IMAvatar [58] shows silhouette artifacts at gracing angles.

plementations of the authors. Note that for IMAvatar, we use the most recent version of the author’s code, which contains additional semantic information for mouth interior and FLAME geometry supervision which is different from the original paper. Figure 4 depicts qualitative results evaluated on the test sequences. To evaluate the image quality of the results quantitatively, we use several pixel-wise metrics; mean squared error, SSIM, PSNR, and the perceptual metric LPIPS [57] (see Table 1). Note that IMAvatar is trained at a resolution of  $256^2$  due to its computational complexity; for the comparison, we upsample the results to  $512^2$ .

All methods produce sharp and photo-realistic images which are hard to distinguish from the ground truth. However, the most noticeable artifacts, especially for the ear regions, were generated by NHA. Moreover, IMAvatar, for some of the videos, had problems with convergence and stability, leading to diverging optimization and premature termination of the training. Compared to these methods,



Figure 4. Qualitative comparisons show that our method produces high-quality facial avatars which beat the state-of-the-art methods in terms of image quality (e.g., capturing fine details like lips and teeth) while being significantly faster to obtain.

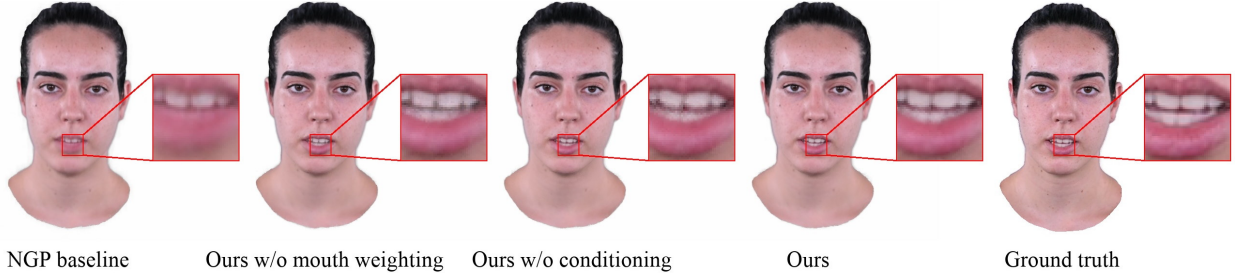


Figure 5. Embedding the neural radiance field around the deformable face model allows us to model dynamic sequences in contrast to the static radiance field of NGP [38]. The expression conditioning and face-parsing-based weighting leads to sharper teeth reconstructions.

Method	L2 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Time ↓
NHA [20]	0.0022	27.71	<b>0.95</b>	<b>0.04</b>	0.63
IMAvatar [58]	0.0023	27.62	0.94	0.06	12.34
NeRFace [16]	<b>0.0018</b>	<b>29.28</b>	<b>0.95</b>	0.07	9.68
Ours	<b>0.0018</b>	28.97	<b>0.95</b>	0.05	<b>0.05</b>

Table 1. Average photometric errors over 19 videos from our dataset, NHA, IMAvatar, and NeRFace datasets (see Fig. 4). The average rendering time of a single frame in seconds is denoted as *Time* in the rightmost column. Our method is on par with NeRFace of Gafni et al. w.r.t. the pixel-wise error metrics. Additionally, our approach achieves low perceptual error in comparison to all methods while being significantly faster to train and evaluate.

our approach can achieve the best image quality while being significantly faster to train (see sup. mat.).

Extrapolation to novel views is an essential aspect of 3D digital avatars that are used in AR or VR applications. In Figure 3, we depict a viewpoint extrapolation comparison with the baseline methods. We can observe that NeRFace [16] produces blurry results in the area of eyes and teeth. IMAvatar [58] exhibits artifacts at gracing angles at the silhouette, and NHA [20] suffers from degenerated geometry with strong artifacts at the ears. In contrast to these methods, our method can robustly generate photo-realistic images under novel poses and achieves high visual quality, especially in the skin and mouth region.

## 5.2. Ablation Studies

We conducted a series of ablation studies to analyze the different components of our pipeline. Specifically, we are interested in the influence of localized expression conditioning for teeth quality (Figure 5), the effect of the geometric prior (Figure 8), especially for the novel view synthesis, and the importance of the deformation field (Figure 6).

**Deformation Field.** Figure 6 shows the impact of the deformation field and the conditioning on the quality of the renderings. We conducted two experiments where we used **a)** a global conditioning instead of the local one and **b)** global conditioning with per-frame learnable codes and without

the deformation field (similar to NeRFace). As can be seen, local conditioning and the mesh-based deformation field helps to avoid overfitting to the short training sequences.

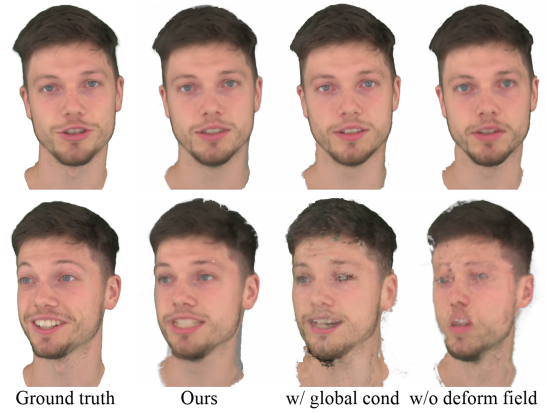


Figure 6. Ablation study w.r.t. the conditioning and deformation field. From left to right: ground truth, ours, ours with global conditioning, and ours without deformation field but with per-frame learnable codes (NeRFace).

**Geometric Prior.** We leverage the geometric prior of the 3DMM FLAME [25] to regularize the depth estimations of our volumetric rendering method. During training, we render depth maps of the per-frame 3DMM reconstructions and measure a loss between the estimated ray termination and the depth of the rendered face model. In Figure 8, we show an ablation study w.r.t. this geometric prior. The generated digital avatar is shown from an unseen profile view, an extreme extrapolation from the training data which observed views in a range of  $\pm 40^\circ$ . Using the additional geometric prior improves the stability and quality of the results.

**Expression Conditioning.** Most publicly available 3DMMs [6, 25] do not explicitly model teeth. However, this region is especially challenging for the reconstruction of 3D facial avatars due to highly dynamic lips, which can occlude the teeth depending on the given expressions. To compen-



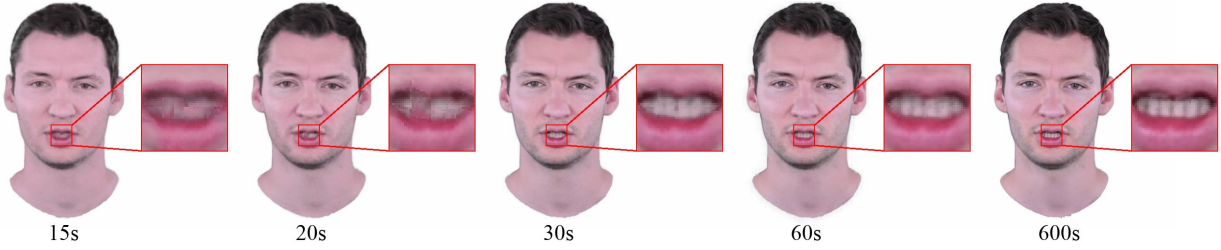


Figure 7. INSTA allows training personalized volumetric avatars from RGB videos within a couple of seconds. Already after 30 seconds of optimization, we achieve good results where the geometry and appearance match the input. To improve the reconstruction of high-frequency details like teeth, the method needs to train approximately 10 min.

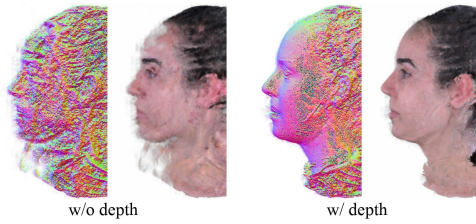


Figure 8. The geometric prior of the 3DMM helps for extrapolation to extreme novel views, in this case,  $90^\circ$ .

sate for the missing geometry, we condition this region on FLAME expression coefficients. In Figure 5, we show that using this additional information helps to improve the synthesis of the mouth interior. Furthermore, we demonstrate that a higher color term weight on the mouth region (Equation (8)) improves the visual quality.

## 6. Discussion

While our method INSTA shows better quality and speed compared to state-of-the-art RGB-video-based avatar generation techniques, there are still several challenges that need to be addressed in future work. Our model handles the dynamically changing facial expressions but does not capture dynamically changing hairs. Thus, the hair quality is not on par with the face interior and still needs improvements in the level of detail. Furthermore, the used 3DMM does not model teeth geometry. A better approximation of the mouth region would increase the viewpoint extrapolation with improved quality of teeth. While our method achieves real-time frame rates for rendering at a resolution of  $512^2$ , the rendering speed needs to be improved to enable high-quality video conferences in AR or VR, especially when a higher resolution is required. With additional engineering, the training process of our method could be moved to a background process that would continuously refine our canonical avatar after an initial warm-up stage. For example, regions initially not visible could be captured during the conversation, and the avatar would be updated accordingly.

## 7. Limitations

An important quality factor of our method is face tracking, as misalignments of the geometry and the images will be propagated to the final avatar. Another limiting aspect is the mouth interior quality due to the lack of geometry in that region, as can be seen in Figure 9.

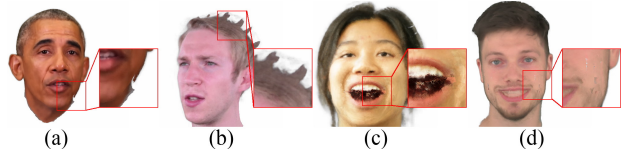


Figure 9. Failure cases: (a) and (b) exhibits outline artifacts at the chin and hair which stem from geometry misalignment of the tracker, (c) extreme expressions can cause artifacts in the mouth region, and (d) extrapolation of expressions can lead to artifacts.

## 8. Conclusion

Instant Volumetric Head Avatars (INSTA) is a novel approach that instantaneously optimizes geometry-guided 3D digital avatars. Our method takes a monocular RGB video as input and optimizes a subject’s dynamic neural radiance field in less than 10 minutes using neural graphics primitives embedded around a 3DMM. In comparisons and ablation studies, we demonstrate the capabilities of INSTA, which enable us to instantaneously create avatars that reflect reality and not a prerecorded appearance that might deviate from the current look of the person. We believe this paradigm change to adaptable online avatars is a stepping stone toward immersive telepresence applications.

**Acknowledgement.** The authors thank all participants of the study and the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting WZ. While TB is a part-time employee of Amazon, his research was performed solely at and exclusively funded by MPI. JT is supported by Microsoft Research gift funds.

## References

- [1] S. Axler, P. Bourdon, and R. Wade. *Harmonic Function Theory*. Graduate Texts in Mathematics. Springer, 2001. 3
- [2] Dejan Azinovic, Ricardo Martin-Brualla, Dan B. Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 6280–6291. IEEE, 2022. 2
- [3] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5835–5844. IEEE, 2021. 2
- [4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5460–5469. IEEE, 2022. 2
- [5] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. volume 23, pages 669–676, 2004. 2
- [6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. pages 187–194, 1999. 7
- [7] Sofien Bouaziz, Yangang Wang, and Mark Pauly. Online modeling for realtime facial animation. *ACM Trans. Graph.*, 32(4):40:1–40:10, 2013. 2
- [8] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhöfer, Shunsuke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shou-I Yu, Yaser Sheikh, and Jason M. Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 41(4):163:1–163:19, 2022. 2
- [9] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. arXiv, 2022. 2
- [10] Wenzheng Chen, Huan Ling, Jun Gao, Edward J. Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9605–9616, 2019. 2
- [11] Xu Chen, Yufeng Zheng, Michael J. Black, Otmar Hilliges, and Andreas Geiger. SNARF: differentiable forward skinning for animating non-rigid neural implicit shapes. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11574–11584. IEEE, 2021. 3, 4
- [12] James H. Clark. Hierarchical geometric models for visible-surface algorithms. page 267, 1976. 2, 4
- [13] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 12872–12881. IEEE, 2022. 2
- [14] Yilun Du, Yanan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 14304–14314. IEEE, 2021. 2
- [15] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5491–5500. IEEE, 2022. 2
- [16] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8649–8658. Computer Vision Foundation / IEEE, 2021. 1, 2, 3, 5, 7
- [17] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. volume abs/2210.06108, 2022. 3
- [18] Stephan J. Garbin, Marek Kowalski, Virginia Estellers, Stanislaw Szymanowicz, Shideh Rezaeifar, Jingjing Shen, Matthew Johnson, and Julien Valentin. Voltemorph: Real-time, controllable and generalisable animation of volumetric representations. arXiv, 2022. 2
- [19] Michael Garland and Paul S. Heckbert. Surface simplification using quadric error metrics. pages 209–216, 1997. 4
- [20] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular RGB videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18632–18643. IEEE, 2022. 1, 2, 5, 7
- [21] Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964. 4
- [22] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. <https://mitsuba-renderer.org>. 2
- [23] Tero Karras. Thinking parallel. <https://developer.nvidia.com/blog/thinking-parallel-part-i-collision-detection-gpu/>, 2012. 4
- [24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5
- [25] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194:1–194:17, 2017. 1, 2, 3, 7

- [26] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 6498–6508. Computer Vision Foundation / IEEE, 2021. 2
- [27] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 3132–3141. IEEE, 2022. 5
- [28] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2
- [29] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.*, 40(6):219:1–219:16, 2021. 2
- [30] Shichen Liu, Weikai Chen, Tianye Li, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7707–7716. IEEE, 2019. 2
- [31] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhöfer, Yaser Sheikh, and Jason M. Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Trans. Graph.*, 40(4):59:1–59:13, 2021. 2
- [32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. volume 34, pages 248:1–248:16, 2015. 2
- [33] Matthew M. Loper and Michael J. Black. Opendr: An approximate differentiable renderer. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII*, volume 8695 of *Lecture Notes in Computer Science*, pages 154–169. Springer, 2014. 2
- [34] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chu-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. volume abs/1906.08172, 2019. 5
- [35] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16169–16178. IEEE, 2022. 2
- [36] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 405–421. Springer, 2020. 2, 3, 4
- [37] Thomas Müller. tiny-cuda-nn, 4 2021. 3, 4
- [38] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multi-resolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 1, 2, 3, 7
- [39] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. pages 5470–5480, 2022. 2
- [40] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5845–5854. IEEE, 2021. 2
- [41] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6):238:1–238:12, 2021. 2
- [42] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Animatable neural radiance fields for human body modeling. volume abs/2105.02872, 2021. 2
- [43] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10318–10327. Computer Vision Foundation / IEEE, 2021. 2
- [44] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. pages 12882–12891, 2022. 2
- [45] Matthew Tancik, Vincent Casser, Xinchun Yan, Sabeek Pradhan, Ben P. Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretschmar. Block-nerf: Scalable large scene neural view synthesis. pages 8238–8248, 2022. 2
- [46] Matthias Teschner, Bruno Heidelberger, Matthias Müller, Danat Pomerantes, and Markus H. Gross. Optimized spatial hashing for collision detection of deformable objects. In Thomas Ertl, editor, *8th International Fall Workshop on Vision, Modeling, and Visualization, VMV 2003, München, Germany, November 19-21, 2003*, pages 47–54. Aka GmbH, 2003. 2
- [47] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason M. Saragih, Matthias Nießner, Rohit Pandey, Sean Ryan Fanello, Gordon Wetstein, Jun-Yan Zhu, Christian Theobalt, Maneesh Agrawala, Eli Shechtman, Dan B. Goldman, and Michael Zollhöfer. State of the art on neural rendering. *Comput. Graph. Forum*, 39(2):701–727, 2020. 2

- [48] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul P. Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Nießner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhöfer, and Vladislav Golyanik. Advances in neural rendering. volume 41, pages 703–735, 2022. [2](#)
- [49] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2387–2395. IEEE Computer Society, 2016. [2](#), [3](#), [5](#)
- [50] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12939–12950. IEEE, 2021. [2](#)
- [51] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd E. Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. pages 5481–5490, 2022. [2](#)
- [52] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5590–5599. IEEE, 2021. [2](#)
- [53] Tianhan Xu, Yasuhiro Fujita, and Eiichi Matsumoto. Surface-aligned neural radiance fields for controllable 3d human synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15862–15871. IEEE, 2022. [2](#)
- [54] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. [3](#)
- [55] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet V2: bilateral network with guided aggregation for real-time semantic segmentation. volume 129, pages 3051–3068, 2021. [4](#), [5](#)
- [56] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. arXiv, 2020. [2](#)
- [57] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. [5](#)
- [58] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M avatar: Implicit morphable head avatars from videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13535–13545. IEEE, 2022. [1](#), [2](#), [3](#), [4](#), [5](#), [7](#)
- [59] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIII*, volume 13673 of *Lecture Notes in Computer Science*, pages 250–269. Springer, 2022. [1](#), [5](#)
- [60] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. *Comput. Graph. Forum*, 37(2):523–550, 2018. [2](#)

## A.3 DRIVABLE 3D GAUSSIAN AVATARS

*Drivable 3D Gaussian Avatars*

Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, Javier Romero.

Published in *International Conference on 3D Vision (3DV)*, Singapore, Republic of Singapore, 2025.

## Abstract

We present Drivable 3D Gaussian Avatars (D3GA), a multi-layered 3D controllable model for human bodies that utilizes 3D Gaussian primitives embedded into tetrahedral cages. The advantage of using cages compared to commonly employed linear blend skinning (LBS) is that primitives like 3D Gaussians are naturally re-oriented and their kernels are stretched via the deformation gradients of the encapsulating tetrahedron. Additional offsets are modeled for the tetrahedron vertices, effectively decoupling the low-dimensional driving poses from the extensive set of primitives to be rendered. This separation is achieved through the localized influence of each tetrahedron on 3D Gaussians, resulting in improved optimization. Using the cage-based deformation model, we introduce a compositional pipeline that decomposes an avatar into layers, such as garments, hands, or faces, improving the modeling of phenomena like garment sliding. These parts can be conditioned on different driving signals, such as key-points for facial expressions or joint-angle vectors for garments and the body. Our experiments on two multi-view datasets with varied body shapes, clothes, and motions show higher-quality results. They surpass PSNR and SSIM metrics of other SOTA methods using the same data while offering greater flexibility and compactness.



# Drivable 3D Gaussian Avatars

Wojciech Zielonka<sup>1,2,3\*</sup>, Timur Bagautdinov<sup>3</sup>, Shunsuke Saito<sup>3</sup>, Michael Zollhöfer<sup>3</sup>,  
Justus Thies<sup>1,2</sup>, Javier Romero<sup>3</sup>

<sup>1</sup>Max Planck Institute for Intelligent Systems, Tübingen, Germany

<sup>2</sup>Technical University of Darmstadt <sup>3</sup>Codec Avatars Lab, Meta

<https://zielon.github.io/d3ga/>



Figure 1. Given a multi-view video input, D3GA is trained to create light, drivable, photorealistic 3D human avatars. These avatars are constructed as a composition of 3D Gaussians encapsulated within tetrahedral cages. The Gaussians undergo transformation and stretching influenced by these cages, are colored using an MLP, and are rasterized into splats. By representing the drivable human as a collection of 3D Gaussian layers, we gain the ability to decompose and manipulate the avatar as needed.

## Abstract

We present *Drivable 3D Gaussian Avatars (D3GA)*, a multi-layered 3D controllable model for human bodies that utilizes 3D Gaussian primitives embedded into tetrahedral cages. The advantage of using cages compared to commonly employed linear blend skinning (LBS) is that primitives like 3D Gaussians are naturally re-oriented and their kernels are stretched via the deformation gradients of the encapsulating tetrahedron. Additional offsets are modeled for the tetrahedron vertices, effectively decoupling the low-dimensional driving poses from the extensive set of primitives to be rendered. This separation is achieved through the localized influence of each tetrahedron on 3D Gaussians, resulting in improved optimization. Using the cage-based deformation model, we introduce a compositional pipeline that decomposes an avatar into layers, such as garments, hands, or faces, improving the modeling of phenomena like garment sliding. These parts can be conditioned on different driving signals, such as keypoints for facial expres-

sions or joint-angle vectors for garments and the body. Our experiments on two multi-view datasets with varied body shapes, clothes, and motions show higher-quality results. They surpass PSNR and SSIM metrics of other SOTA methods using the same data while offering greater flexibility and compactness.

## 1. Introduction

Developing drivable, photorealistic human avatars is crucial for better long-distance telecommunication that provides an immersive experience to the users. The motion and deformations across various segments of a complex avatar’s body are influenced by distinct signals, such as facial expressions and body movements. This complexity poses challenges for accurate modeling using a single layer. Multi-layered avatars become essential to represent these different regions, ensuring proper motion and visual fidelity. Similarly, garments present challenges such as sliding, necessitating separate modeling of each clothing piece.

Mixture of Volumetric Primitives (MVP) [30] started a

\*Work done while Wojciech Zielonka was an intern at the Codec Avatars Lab in Pittsburgh, PA, USA

successful line of hybrid implementations, where volumetric primitives are embedded on the surface of the tracked mesh. This representation, despite excellent results, struggles when the provided mesh is not precise or lacks details, ultimately producing artifacts and misaligning the primitives. Similar CNN-based architectures [1, 27, 29, 30, 52], do not allow for easy garment decomposition and assume a fixed amount of 3D primitives since the CNN size has to be set for the training. Furthermore, numerous methods [1, 24, 30, 57] lack the capability of layered conditioning specific to different body parts. For example, they may not support using keypoints for the face or motion vectors for clothing like t-shirts. This is an important aspect of a holistic system that, ultimately, needs to capture speech, face, gestures, and garment motion. State-of-the-art drivable avatars [52, 68] require dense input signals like RGB-D images or even multi-view camera setups at test time, which might not be suitable for low-bandwidth connections in telepresence applications. Finally, drivable NeRFs and 3DGS avatars typically rely on LBS to transform samples between canonical and observation spaces. However, LBS is limited by the low degree of freedom of the model, whereas cages can handle more complex non-linear motion and offer additional physical properties (e.g., stretching).

We designed our method to use a minimal set of inputs and still be competitive with the ones that require more information to train an avatar. D3GA models digital humans using volumetric primitives represented as 3D Gaussians embedded into a tetrahedral cage which is naturally described by phenomenons like stretching, rotation, and scaling. Accordingly, instead of LBS, our method builds on a classic deformation model for transforming volumes [40]. Specifically, by recasting cages from the canonical space into a deformed one, the 3D Gaussian covariance matrices undergo the encapsulating tetrahedral deformation transformation. Recent advancements in incorporating physics into Gaussians [8, 70] show further promise in the context of cage usage for garment modeling by capitalizing on [4, 35]. Also, cages decouple the representation resolution (related to the amount of Gaussians) from the degrees of freedom present in the model ultimately allowing an effective regularization of the deformations in contrast to LBS which depends on the global bone transformations only. In addition, we employ a compositional structure based on separate body, face, and garment cages, allowing us to model those parts independently, including localized conditioning based on different driving signals (e.g., keypoints).

We train person-specific models on nine high-quality multi-view sequences with a wide range of body shapes, motion, and clothing (not limited to tight-fitting), which later can be driven with new poses from any subject.

In summary, we present Drivable 3D Gaussian Avatars (D3GA) with the following contributions:

- A light, flexible, and composable model based on 3D Gaussian primitives driven by tetrahedral cage-based deformations which improve their body modeling properties.
- Localized motion conditioning which enables for instance facial expressions.

## 2. Related Work

D3GA reconstructs controllable digital full-body avatars using multi-view video and joint angle motion by combining 3D Gaussian Splatting (3DGS) [19] with cage-based deformations [12, 14, 17]. Current methods for controllable avatars rely on dynamic Neural Radiance Fields (NeRF) [38, 43, 44], point-based [34, 71, 77], or hybrid representations [1, 6, 30, 79], which are either slow to render or fail to correctly disentangle garments from the body, leading to poor generalization to new poses. Recently, incorporating 3DGS into dynamic scenarios has opened new research avenues [27, 49, 69, 72, 76]. For a thorough overview, we refer readers to state-of-the-art reports on digital avatars and neural rendering [60, 61, 82].

**Dynamic Neural Radiance Fields** NeRF [39] is a popular appearance model for human avatars, representing scenes volumetrically with density and color information using an MLP. Images are rendered via ray casting and volumetric integration of sample points [18]. Many methods have successfully applied NeRF to dynamic scenes [9, 26, 43, 44, 47, 65, 71, 79], achieving high-quality results. However, most methods treat avatars as a single layer [24, 38, 45, 55–57, 78], which complicates modeling phenomena like sliding or loose garments. Methods like [6, 7] address this using a hybrid representation, combining explicit geometry from SMPL[31] with implicit dynamic NeRF. Despite impressive garment reconstruction, these methods struggle with novel pose prediction. TECA [74] extends SCARF to a generative framework, enabling prompt-based generation of NeRF-based accessories and hairstyles.

**Point-based Rendering** Before 3DGS, many methods used point-based rendering [34, 57, 77] or sphere splatting [23], with optimizable positions and sizes. NPC by Su et al. [57] defines a point-based body model for avatar representation, but requires lengthy nearest neighbor searches during training (12 hours vs. 30 minutes for our model), making it impractical for dense multi-view datasets. Ma et al. [34] represent garments as a pose-dependent function mapping SMPL points [31] to the clothing space. This is improved in [48] with a neural deformation field, but both models only address geometry, not appearance. Zheng et al. [77] represent the upper part of an avatar as a point cloud, grown during optimization and rasterized using a differentiable renderer [63]. While achieving photorealistic local results, the avatars suffer from artifacts like holes.

**Cage-based Deformations** Cages[40] are commonly used for geometry modeling and animation, serving as sparse proxies to control all interior points, enabling efficient deformation by manipulating only cage nodes. Yifan et al. [64] introduced neural cages for detail-preserving shape deformation, where a neural network rigs the source object into the target via a proxy. Garbin et al. [10] extended dynamic NeRF with tetrahedron cages to unposed ray samples based on tetrahedron intersections. This method is real-time, high-quality, and controllable, but limited to objects with local deformations like heads, and not suitable for highly articulate objects like full-body avatars. Peng et al. used a cage to deform a radiance field in CageNeRF [46]. While their low-resolution cages can be applied to full-body avatars, they fail to model detailed features like faces or complex deformations.

**Time-conditioned Methods** Playback methods [2, 5, 13, 25, 66, 73] represent a scene as a time-conditioned function that cannot be arbitrarily controlled, allowing only for a novel viewpoint synthesis while traversing the time axis. Yang et al. [73] extended the representation of 3DGS [19] into 4DGS, effectively incorporating time into the primitive representation. Wu et al. [66] combine Gaussians with 4D neural voxels, inspired by HexPlane [2], which achieves real-time rendering and novel-view synthesis. However, these solutions fall into a different class of algorithms compared to pose-conditioned drivable avatars, which is our goal.

**Dynamic Gaussian Splatting** D3GA is based on 3D Gaussian Splatting (3DGS) [19], a recent alternative to NeRF for modeling neural scenes. Due to its real-time capabilities and high-quality results, 3DGS has inspired numerous follow-up papers [8, 15, 33, 49, 69, 70, 72, 76, 80, 81] in areas such as physics simulation, hair modeling, head avatars, and fluid dynamics. Several works [27, 41, 53] recently introduced convolutional networks to regress Gaussian maps. Despite achieving high-quality results, fixed convolutional architectures do not allow for local conditioning or adjusting the number of Gaussians during training. These methods also use up to 23 times more parameters, causing the model size to reach almost 1 GiB. In contrast, our pipeline remains lightweight and flexible, offering garment decomposition and localized conditioning. Finally, using CNNs can slow down the pipeline to around 10 FPS [27], whereas our method remains real-time.

### 3. Method

D3GA is built on 3DGS extended by a neural representation and tetrahedral cages to model the color and geometry of each dynamic part of the avatar, respectively. In the following, we introduce the formulation of 3D Gaussian Splatting and give a detailed description of our method.

#### 3.1. 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) [19] is designed for real-time novel view synthesis in multi-view static scenes. Their rendering primitives are scaled 3D Gaussians [22, 63] with a 3D covariance matrix  $\Sigma$  and mean  $\mu$ :

$$G(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)}. \quad (1)$$

To splat the Gaussians, Zwicker et al. [83] define the projection of 3D Gaussians onto the image plane as:

$$\Sigma' = \mathbf{A} \mathbf{W} \Sigma \mathbf{W}^T \mathbf{A}^T, \quad (2)$$

where  $\Sigma'$  is a covariance matrix in 2D space,  $\mathbf{W}$  is the view transformation, and  $\mathbf{A}$  is the Jacobian of the affine approximation of the projective transformation. During optimization, enforcing the positive semi-definiteness of the covariance matrix  $\Sigma$  is challenging. To avoid this, Kerbl et al. [19] use an equivalent formulation of a 3D Gaussian as a 3D ellipsoid parameterized with a scale  $\mathbf{S}$  and rotation  $\mathbf{R}$ :

$$\Sigma = \mathbf{R} \mathbf{S} \mathbf{S}^T \mathbf{R}^T. \quad (3)$$

3DGS uses spherical harmonics [51] to model the view-dependent color of each Gaussian. In practice, appearance is modeled with an optimizable 48 elements vector representing four bands of spherical harmonics.

#### 3.2. Body Cage Creation

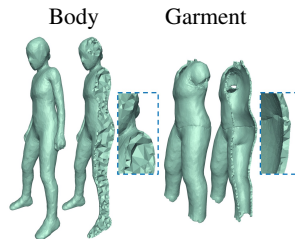


Figure 3. D3GA uses a tetrahedral mesh for deformation modeling.

To deform 3D Gaussians, we utilize tetrahedron cage-based deformations as a coarse proxy for the body, face, and individual garments. Unlike a triangle, which is two-dimensional,

a tetrahedron is a polyhedron with four triangular faces (A, B, C, D), providing a three-dimensional structure. The volume of a tetrahedron can be calculated using the scalar triple product of vectors, which enables precise control and deformation of the enclosed 3D Gaussians. The volume  $V$  is given by:

$$V = \frac{1}{6} |\mathbf{AB} \cdot (\mathbf{AC} \times \mathbf{AD})| \quad (4)$$

where  $\mathbf{AB}$ ,  $\mathbf{AC}$ ,  $\mathbf{AD}$  are edges of tetrahedron. This property allows us to compute the deformation gradient similarly to Sumner et al. [58] and transfer it to the Gaussian covariance matrix (Equation 7), see Supp. Mat, for more details.

To create a cage per garment, we segment all images of a single time instance using an EfficientNet [59] backbone

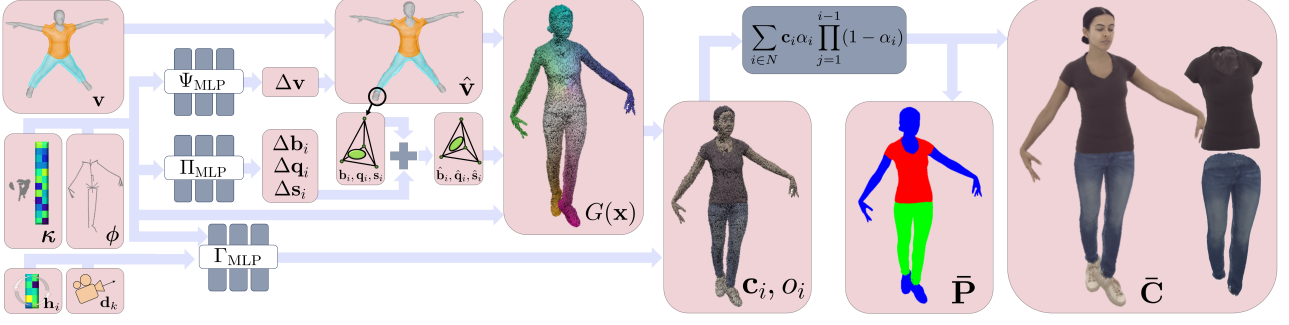


Figure 2. **Overview.** D3GA uses 3D pose  $\phi$ , face embedding  $\kappa$ , viewpoint  $\mathbf{d}_k$  and canonical cage  $\mathbf{v}$  (as well as auto-decoded color features  $\mathbf{h}_i$ ) to generate the final render  $\bar{\mathbf{C}}$  and auxiliary segmentation render  $\bar{\mathbf{P}}$ . The inputs in the left are processed through three networks ( $\Psi_{\text{MLP}}$ ,  $\Pi_{\text{MLP}}$ ,  $\Gamma_{\text{MLP}}$ ) per avatar part to generate cage displacements  $\Delta\mathbf{v}$ , Gaussians deformations  $\mathbf{b}_i, \mathbf{q}_i, \mathbf{s}_i$  and color/opacity  $\mathbf{c}_i, o_i$  respectively. After cage deformations transform canonical Gaussians, they are rasterized into the final images according to Eq. 10.

with PointRender [21] refinement, trained on a corpus of similar multi-view captures. The per-image 2D segmentation masks are projected onto a body mesh  $\hat{\mathbf{M}}$  to obtain per-triangle labels (body, upper, lower). To get the mesh  $\hat{\mathbf{M}}$ , we fit a low-resolution LBS model to a single 3D scan of the subject and then fit such model to the segmented frame by minimizing the distance to the 3D keypoints, extracted with an EfficientNet trained on similar captures. We transform the body mesh into canonical space with LBS and divide it into body part templates  $\mathbf{M}_k$ . The garment meshes are additionally inflated by 1-3 cm along the vertex normals. Afterward, we run a voxelization of the meshes and subsequently extract the mesh using the marching cubes algorithm [32]. After that, we use TetGen [54] to turn the unposed meshes  $\mathbf{M}_k$  into tetrahedral meshes  $\mathbf{T}_k$ . Consequently, cages for garments are hollow, containing only their outer layer, while the body cage is solid (Figure 3). The face cage is composed of the body tetrahedra which contains triangles defined as the face on the LBS template. The cage nodes are deformed according to LBS weights transferred from the closest vertex in  $\mathbf{M}_k$ .

### 3.3. Cage Deformation Transfer

While classic cage methods typically deform the volume according to complex weight definitions [14, 16, 17], using linear weights works well in practice when cage cells are small, making it easier to integrate into an end-to-end training system. Specifically, we define  $\mathbf{v}_{ij}$  as the vertices of tetrahedron  $i$  in canonical space, any point  $\mathbf{x}$  inside this tetrahedron can be defined by its barycentric coordinates  $b_j$ :

$$\mathbf{x} = \sum_{j=1}^4 b_j \mathbf{v}_{ij}. \quad (5)$$

Each Gaussian 3D mean  $\boldsymbol{\mu} = \mathbf{x}$  is obtained as a linear combination of learnable barycentric coordinates  $b_j$  and tetrahedron vertices  $\mathbf{v}_{ij}$ . When the tetrahedra are transformed to posed space according to  $\hat{\mathbf{v}}_{ij} = \text{LBS}(\mathbf{v}_{ij}, \phi, \mathbf{w}_{ij})$ , where  $\phi$  is the pose and  $\mathbf{w}_{ij}$  are the blendweights, the same linear relation holds  $\hat{\mathbf{x}} = \sum_{j=1}^4 b_j \hat{\mathbf{v}}_{ij}$ . To leverage the cage

volume properties (rotation, sheer, and scaling), we use the deformation gradient [58]:

$$\mathbf{J}_i \mathbf{E}_i = \hat{\mathbf{E}}_i, \quad (6)$$

$$\mathbf{J}_i = \hat{\mathbf{E}}_i \mathbf{E}_i^{-1}, \quad (7)$$

where  $\hat{\mathbf{E}}_i \in \mathbb{R}^{3 \times 3}$  and  $\mathbf{E}_i \in \mathbb{R}^{3 \times 3}$  contain three edges from tetrahedron  $i$  defined in deformed and canonical spaces, respectively. The gradient  $\mathbf{J}_i$  is used to transform the kernel of each Gaussian  $i$  (Eq 8). See Supp. mat for more details.

### 3.4. Drivable Gaussian Avatars

We initialize a fixed number of Gaussians, whose 3D means  $\boldsymbol{\mu}$  are sampled on the surface of  $\hat{\mathbf{M}}$ . However, we are not limited to the fixed amount of Gaussians allowing for cloning or densification if needed. The rotation of each Gaussian is initialized so that the first two axes are aligned with the triangle surface and the third one with the normal: this is a good approximation for a smooth surface. The scale is initialized uniformly across a heuristic range depending on inter-point distances as in [19]. We assign each sampled position  $\mathbf{x}$  to the intersecting tetrahedron and compute its barycentric coordinates  $\mathbf{b} \in \mathbb{R}^4$ . To deform the tetrahedron volume, we incorporate the deformation gradient  $\mathbf{J}$  defined in Eq. 7 into the Gaussian covariance matrix from Eq. 3.

This is an important step as the deformation gradient  $\mathbf{J}$  encapsulates many phenomena that we want to model, for instance, rotation, stretching, and sheering. To correctly transfer the deformation to 3D Gaussian primitives, we apply it to the covariance matrix  $\boldsymbol{\Sigma}$ , effectively modeling the 3DGS ellipsoids depending on the shape deformation from the canonical space into deformed one. Thus, the final covariance matrix passed to the rasterizer is denoted as:

$$\hat{\boldsymbol{\Sigma}} = \mathbf{J}_i \boldsymbol{\Sigma} \mathbf{J}_i^T, \quad (8)$$

where  $\mathbf{J}_i$  is the deformation gradient of the tetrahedron containing the 3D mean of the Gaussian with covariance  $\boldsymbol{\Sigma}$ . This way, we transfer the deformation into the Gaussians, improving modeling phenomena like garment stretching.



Each part of the avatar (the garment, body, or face) is controlled by a separate GaussianNet  $\mathbb{G}_{\text{Net}} = \{\Gamma_{\text{MLP}}, \Pi_{\text{MLP}}, \Psi_{\text{MLP}}\}$  which is defined as a set of small specialized multi-layer perceptrons (MLP) parametrized as:

$$\begin{aligned}\Psi_{\text{MLP}} &: \{\phi, \text{enc}_{\text{pos}}(\mathbf{v})\} \rightarrow \Delta \mathbf{v}, \\ \Pi_{\text{MLP}} &: \{\phi, \mathbf{b}_i, \mathbf{q}_i, \mathbf{s}_i\} \rightarrow \{\Delta \mathbf{b}_i, \Delta \mathbf{s}_i, \Delta \mathbf{q}_i\}, \\ \Gamma_{\text{MLP}} &: \{\phi, \text{enc}_{\text{view}}(\mathbf{d}_k), \mathbf{h}_i, \mathbf{f}_j\} \rightarrow \{\mathbf{c}_i, o_i\}.\end{aligned}\quad (9)$$

All the networks take joint angles  $\phi$  (or face encodings  $\kappa$  for the face networks) as inputs, in addition to network-specific conditioning. The cage node correction network  $\Psi_{\text{MLP}}$  takes positional encodings [39] for all canonical vertices to transform them into offsets of the cage node positions similar to SMPL [31] pose-correctives. To adapt our representation further to the pose, the Gaussian correction network  $\Pi_{\text{MLP}}$  takes the canonical Gaussian parameters (barycentric coordinates  $\mathbf{b}_i \in \mathbb{R}^4$ , rotation  $\mathbf{q}_i \in \mathbb{R}^4$  and scale  $\mathbf{s}_i \in \mathbb{R}^3$ ) to predict corrections of those same parameters. These two networks are necessary to capture high-frequency details outside the parametric transformation.

The shading network  $\Gamma_{\text{MLP}}$  transforms encoded view direction and initial color into final color and opacity,  $\mathbf{c}_i, o_i$ . Unlike 3DGS, we use a pose-dependent color representation to model self-shadows and wrinkles in garments. The view angle is projected onto the first four spherical harmonics bands  $\text{enc}_{\text{pos}}(\cdot)$ , while the initial color is an auto-decoded feature vector  $\mathbf{h}_i$  [42]. Additionally, the face region utilizes face embeddings  $\kappa$  as input instead of pose  $\phi$ . This adaptability stems from our model’s composability and holds the potential for extension to other regions, such as hair, shoes, or hands. A small auxiliary MLP regresses  $\kappa$  based on 150 3D keypoints  $\mathbf{k}$  normalized by their training mean and standard deviations. This effectively enables us to model facial expressions.

Finally, we also add an embedding vector with the time frame of the current sample [36]. This allows D3GA to explain properties that cannot be modeled (e.g., cloth dynamics) from our input, effectively avoiding excessive blur due to averaging residuals. During testing, the average training embedding is used.

### 3.5. Training Objectives

As in 3DGS [19], we define the color  $\bar{\mathbf{C}}$  of pixel  $(u, v)$ :

$$\bar{\mathbf{C}}_{u,v} = \sum_{i \in \mathcal{N}} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (10)$$

where  $\mathbf{c}_i$  is the color predicted by  $\Gamma_{\text{MLP}}$ , which replaces the spherical harmonics in 3DGS.  $\alpha_i$  is computed as the product of the Gaussian density in Eq. 1 with covariance matrix  $\Sigma'$  from Eq. 2 and the learned per-point opacity  $o_i$  predicted by  $\Gamma_{\text{MLP}}$ . The sum is computed over set  $\mathcal{N}$ , the Gaussians

with spatial support on  $(u, v)$ . The primary loss in D3GA is a weighted sum of three different color losses applied to the estimated image  $\bar{\mathbf{C}}$  and the ground truth RGB image  $\mathbf{C}$ :

$$\mathcal{L}_{\text{Color}} = (1 - \omega) \mathcal{L}_1 + \omega \mathcal{L}_{\text{D-SSIM}} + \zeta \mathcal{L}_{\text{VGG}}, \quad (11)$$

where  $\omega = 0.2$ ,  $\zeta = 0.005$  (after 400k iterations steps and zero otherwise),  $\mathcal{L}_{\text{D-SSIM}}$  is a structural dissimilarity loss, and  $\mathcal{L}_{\text{VGG}}$  is the perceptual VGG loss.

To encourage correct garment separation, we introduce a garment loss. Since each Gaussian  $i$  is statically assigned to a part, we define  $\mathbf{p}_i$  as a constant-per-part color and consequently render  $\bar{\mathbf{P}}$  by replacing  $\mathbf{c}_i$  by  $\mathbf{p}_i$  in Eq. 10. Then, we compute the  $\mathcal{L}_1$  norm between predicted parts  $\bar{\mathbf{P}}$  and ground truth segmentations  $\mathbf{P}$ ,  $\mathcal{L}_{\text{Garment}} = \mathcal{L}_1(\bar{\mathbf{P}}, \mathbf{P})$ . Moreover, we are using the Neo-Hookean loss based on Macklin et al. [35] to enforce the regularization of the predicted tetrahedra for the regions with low supervision signal:

$$\mathcal{L}_{\text{Neo}} = \frac{1}{N} \sum_{i=0}^N \frac{\lambda}{2} (\det(\mathbf{J}_i) - 1)^2 + \frac{\mu}{2} (\text{tr}(\mathbf{J}_i^T \mathbf{J}_i) - 3), \quad (12)$$

where  $\mathbf{J}_i$  denotes the deformation gradient between a canonical and a deformed tetrahedron (Eq. 7),  $N$  is the total number of tetrahedrons, and  $\lambda$  and  $\mu$  are the Lamé parameters [35]. The overall loss is defined as:

$$\mathcal{L} = \nu \mathcal{L}_{\text{Color}} + \nu \mathcal{L}_{\text{Garment}} + \tau \mathcal{L}_{\text{Neo}}, \quad (13)$$

where  $\nu = 10$  and  $\tau = 0.005$  balance the different losses.

We implemented D3GA based on the differentiable 3DGS renderer [19]. The networks  $\Pi_{\text{MLP}}, \Psi_{\text{MLP}}, \Gamma_{\text{MLP}}$  have three hidden layers with 128 neurons and ReLU activation functions. In our experiments, we train the networks for 700k (Ours) and 400k (ActorsHQ) steps with a multi-step scheduler with a decay rate of 0.33, a batch size of one, and using the Adam optimizer [20] with a learning rate set to  $5e - 4$ . We ran all experiments on a single Nvidia V100 GPU with  $1024 \times 667$  images. When ground truth poses are not available, as in the case of ActorsHQ [13], we additionally refine poses regressed from keypoints during avatar training and during the test time, and optionally projected them onto PCA basis computed from the training set.

## 4. Dataset

Our dataset comprises nine subjects performing various motions, observed by 200 cameras. We use 12,000 frames for training (at 10 FPS) and 1,500 for testing (at 30 FPS). Images were captured at a resolution of  $4096 \times 2668$  in a multi-view studio with synchronized cameras and downsampled to  $1024 \times 667$  to reduce computational cost. We utilize 2D segmentation masks, RGB images, keypoints, and 3D joint angles for training, as well as a single registered mesh to create our template  $\hat{\mathbf{M}}$ . Of the nine subjects, data for four



Figure 4. Qualitative comparisons show that D3GA models facial expressions and garments better than other SOTA approaches. Especially regions with loose garments like skirts or sweatpants.

is publicly available through the Goliath-4 dataset release [37].

## 5. Results

We evaluate and benchmark our method w.r.t. five state-of-the-art multiview-based solutions [1, 11, 27, 30, 50]. We compare D3GA to the mesh-based full-body avatar methods BodyDecoder (BD) [1] and MVP-based avatars [30, 52] evaluated on our dataset.

Additionally, we evaluated D3GA on the ActorsHQ dataset [13] using a significantly smaller number of cameras (40). We compare to SOTA pose-conditioned 3DGS avatar methods, including Animatable Gaussians (AG) [27], 3DGS-Avatar [50], and Gaussian Avatar (GA) [11] which were trained on the same multiview data.

Please note that our method, along with 3DGS-Avatar and GA, represents a lightweight class of MLP-based algorithms, utilizing up to 10 million parameters. In contrast, the CNN-based MVP, BD, and AG [27] which in this case uses approximately 23 times more parameters (230 million).

### 5.1. Image Quality Evaluation

Our model is evaluated using SSIM, PSNR, and the perceptual metric LPIPS [75], with random color backgrounds. For the ActorsHQ evaluation, we utilized SMPL-X fittings

obtained through OpenPose [3] and scan-to-mesh optimization. Table 1 shows that our method achieves the best PSNR and SSIM on our dataset compared to MVP [30] and BD [1]. Furthermore, on the ActorsHQ dataset, D3GA outperforms other Gaussian Avatar methods in terms of PSNR and SSIM. However, similar to previous evaluations, our method lacks sharpness due to its much smaller size compared to the CNN-based architecture of AG [27]. Moreover, our approach allows us to decompose avatars into drivable layers, unlike other volumetric methods. Each separate garment layer can be controlled solely by skeleton joint angles, without requiring specific garment registration modules as in [67].

Dataset	Method	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$
Ours	Ours	<b>30.634</b>	0.054	<b>0.964</b>
	MVP [30]	28.795	0.051	0.955
	BD [1]	29.918	<b>0.044</b>	0.959
ActorsHQ	Ours	<b>26.562</b>	0.065	<b>0.944</b>
	GA [11]	24.731	0.088	0.933
	3DGS-Avatar [50]	21.709	0.082	0.915
	AG [28]	26.454	<b>0.055</b>	0.937

Table 1. Our method scores the best in terms of PSNR and SSIM compared to BD [1] and MVP [30] on our dataset. D3GA is the best among MLP-based avatars, ranking only second in terms of sharpness compared to AG, which uses a CNN-based architecture.



Figure 5. ActorsHQ [13] comprises challenging garments that contain high-frequency patterns. Our method despite its small size can capture it and performs the best in terms of PSNR and SSIM, ranking second only in terms of sharpness to AG [27], which presents very sharp results due to the powerful StyleUNet [62].

## 5.2. Ablation Studies

**Importance of cage deformations** We replaced tetrahedrons with triangles to emphasize the crucial role of cage deformation gradients in transforming Gaussians. We modified Eq 5 such that 3D means are obtained through the barycentric coordinates of triangles  $\mathbf{b} \in \mathbb{R}^3$  instead of tetrahedrons  $\mathbf{b} \in \mathbb{R}^4$ . The rest of the pipeline remains unchanged, with MLPs computing the same corrective terms as our cage-based model. Since triangles do not provide volume, we disabled the application of the cage deformation

gradient  $\mathbf{J}$ , but the Gaussians are still modeled by the predicted residuals w.r.t. the canonical space. Figure 8 shows that the triangle-based approach does not stretch the primitives correctly, creating holes and artifacts which demonstrates the importance of using cages for deformation.

**Garment loss** The garment loss  $\mathcal{L}_{Garment}$  (Fig. 7) serves two primary purposes: it improves garment separation and reduces erroneously translucent regions. We can observe qualitatively that regions between garments’ boundaries without the regularizer are blurry and have erroneous opac-





Figure 6. D3GA enables motion transfer showing good generalizability while preserving each avatar’s high-quality details.

Method	#parameters (M)	size (MiB)
Ours	9	45
GaussianAvatar [11]	7	59
3DGS-Avatar [50]	6	57
AG [28]	232	862

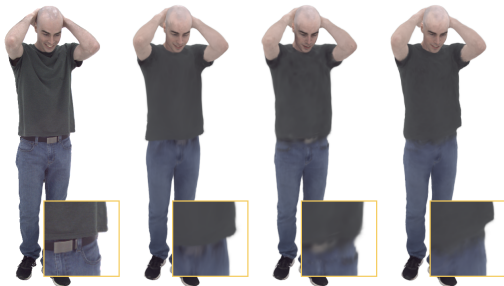
Table 2. Model compactness. D3GA offers the best tradeoff between quality and model size.

ity, see supp. mat. **Single layer avatar** D3GA supports a single-layer training for the garment and body, which struggles to model proper garment sliding. The results are presented in the last column of Fig. 7. It can be observed that the edges between the T-shirt and jeans are over-smoothed.

**Size and compactness** Our model offers an optimal balance between quality and model size, making it both compact and easily portable. This lightweight representation sets D3GA apart from much larger and more cumbersome models like AG [27]. As shown in Table 2, D3GA is similar in size to other methods, yet it delivers superior quality compared to models in the same category. This makes D3GA an attractive choice for telepresence applications, where both efficiency and performance are crucial.

## 6. Discussion

While D3GA shows better quality and competitive rendering speed w.r.t. the state of the art, there are still particular challenges. High-frequency patterns, like stripes, may result in blurry regions. One way of improving image quality would be using a variational autoencoder to regress Gaussian parameters per texel of a guide mesh similar to [27, 30].



Ground Truth   Ours   w/o  $\mathcal{L}_{Garment}$    Single Layer

Figure 7. Ablation of D3GA: shape smoothness without  $\mathcal{L}_{Garment}$ , and sliding artifacts with a single layer representation.

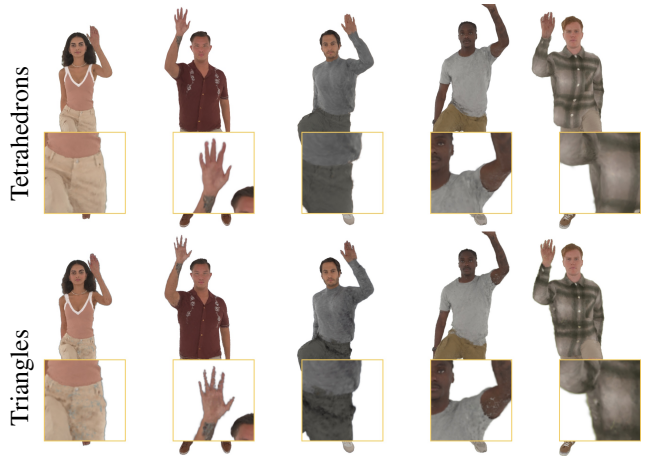


Figure 8. Gaussian primitives embedded in triangles, compared to tetrahedrons, produce more artifacts, resulting in small holes and reduced sharpness that is reflected in the LIPIS score, which drops from 0.0648 to 0.0703.

Despite using the  $\mathcal{L}_{Garment}$  loss, self-collisions for loose garments are still challenging, and the sparse controlling signal does not contain enough information about complex wrinkles or self-shadowing. A potential solution to solve self-penetration would be to incorporate explicit collision detection [4] for the tetrahedrons. An exciting follow-up work direction would be replacing the appearance model in D3GA with a relightable one. D3GA is currently limited to model photorealistic avatars for a few consenting subjects captured in a dense multi-view capture device. While this limits the potential misuse of the technology of driving somebody else’s avatar without their consent, it needs to be addressed in future work. In conclusion, it’s worth noting that the D3GA offers significant flexibility and can be customized for particular applications. For instance, one could employ additional Gaussians to capture high-frequency detail or opt to eliminate garment supervision, particularly if precise cage geometry decomposition isn’t necessary.

## 7. Conclusion

We have proposed D3GA, a novel approach for reconstructing multi-layered animatable human avatars using tetrahedral cages embedded with 3D Gaussians. To transform the rendering primitives from canonical to deformed space, we directly apply the deformation gradient to the 3D Gaussian parametrization, enabling improved avatar modeling. Our method’s compositional approach enables various forms of localized conditioning, such as using keypoints for facial expressions, and can be extended to other regions like hair, hands, or shoes. This capability is essential for creating holistic avatars driven by diverse input signals. We have demonstrated high-quality results that surpass state-of-the-art methods with similar model architectures, all while maintaining a lightweight, real-time, and compact approach.



## References

- [1] Timur M. Bagautdinov, Chenglei Wu, Tomas Simon, Fabián Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason M. Saragih. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)*, 40:1 – 17, 2021. [2](#), [6](#)
- [2] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [3](#)
- [3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. [6](#)
- [4] Heng Chen, Elier Diaz, and Cem Yuksel. Shortest path to boundary for self-intersecting meshes. *ACM Transactions on Graphics (TOG)*, 42:1 – 15, 2023. [2](#), [8](#)
- [5] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. [3](#)
- [6] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. *SIGGRAPH Asia 2022 Conference Papers*, 2022. [2](#)
- [7] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J. Black. Learning disentangled avatars with hybrid 3d representations. *arXiv*, 2023. [2](#)
- [8] Yutao Feng, Xiang Feng, Yintong Shang, Ying Jiang, Chang Yu, Zeshun Zong, Tianjia Shao, Hongzhi Wu, Kun Zhou, Chenfanfu Jiang, and Yin Yang. Gaussian splashing: Dynamic fluid synthesis with gaussian splatting. *ArXiv*, abs/2401.15318, 2024. [2](#), [3](#)
- [9] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8645–8654, 2020. [2](#)
- [10] Stephan J. Garbin, Marek Kowalski, Virginia Estellers, Stanislaw Szymanowicz, Shideh Rezaeifar, Jingjing Shen, Matthew Johnson, and Julien Valentin. Voltemorph: Real-time, controllable and generalisable animation of volumetric representations. *CoRR*, abs/2208.00949, 2022. [3](#)
- [11] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. [6](#), [7](#), [8](#)
- [12] Jin Huang, Xiaohan Shi, Xinguo Liu, Kun Zhou, Li-Yi Wei, Shang-Hua Teng, Hujun Bao, Baining Guo, and Harry Shum. Subspace gradient domain mesh deformation. *ACM SIGGRAPH 2006 Papers*, 2006. [2](#)
- [13] Mustafa Isik, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Trans. Graph.*, 42(4):160:1–160:12, 2023. [3](#), [5](#), [6](#), [7](#)
- [14] Alec Jacobson, Ilya Baran, Jovan Popović, and Olga Sorkine-Hornung. Bounded biharmonic weights for real-time deformation. *ACM SIGGRAPH 2011 papers*, 2011. [2](#), [4](#)
- [15] Ying Jiang, Chang Yu, Tianyi Xie, Xuan Li, Yutao Feng, Huamin Wang, Minchen Li, Henry Lau, Feng Gao, Yin Yang, and Chenfanfu Jiang. Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. *ArXiv*, abs/2401.16663, 2024. [3](#)
- [16] Pushkar Joshi, Mark Meyer, Tony DeRose, Brian Green, and Tom Sanocki. Harmonic coordinates for character articulation. *ACM Trans. Graph.*, 26(3):71, 2007. [4](#)
- [17] Tao Ju, Scott Schaefer, and Joe D. Warren. Mean value coordinates for closed triangular meshes. *ACM SIGGRAPH 2005 Papers*, 2005. [2](#), [4](#)
- [18] James T. Kajiya. The rendering equation. *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, 1986. [2](#)
- [19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42:1 – 14, 2023. [2](#), [3](#), [4](#), [5](#)
- [20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. [5](#)
- [21] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick. Pointrend: Image segmentation as rendering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9796–9805. Computer Vision Foundation / IEEE, 2020. [4](#)
- [22] Georgios Kopanas, Julien Philip, Thomas Leimkühler, and George Drettakis. Point-based neural rendering with per-view optimization. *Computer Graphics Forum*, 40, 2021. [3](#)
- [23] Christoph Lassner and Michael Zollhöfer. Pulsar: Efficient sphere-based neural rendering. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1440–1449, 2021. [2](#)
- [24] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. *ArXiv*, abs/2206.08929, 2022. [2](#)
- [25] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, and Zhaoyang Lv. Neural 3d video synthesis from multi-view video. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. [3](#)
- [26] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6494–6504, 2020. [2](#)
- [27] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian

- maps for high-fidelity human avatar modeling. *ArXiv*, abs/2311.16096, 2023. 2, 3, 6, 7, 8
- [28] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6, 8
- [29] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor. *ACM Transactions on Graphics (TOG)*, 40:1 – 16, 2021. 2
- [30] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason M. Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (TOG)*, 40:1 – 13, 2021. 1, 2, 6, 8
- [31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2015. 2, 5
- [32] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *SIGGRAPH '87: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, pages 163–169, 1987. 4
- [33] Haimin Luo, Ouyang Min, Zijun Zhao, Suyi Jiang, Longwen Zhang, Qixuan Zhang, Wei Yang, Lan Xu, and Jingyi Yu. Gaussianhair: Hair modeling and rendering with light-aware gaussians. *ArXiv*, abs/2402.10483, 2024. 3
- [34] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10954–10964, 2021. 2
- [35] Miles Macklin and Matthias Müller. A constraint-based formulation of stable neo-hookean materials. *Proceedings of the 14th ACM SIGGRAPH Conference on Motion, Interaction and Games*, 2021. 2, 5
- [36] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7206–7215, 2020. 5
- [37] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venstain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Daur, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024. 6
- [38] Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. Keypointnerf: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. *ArXiv*, abs/2205.04992, 2022. 2
- [39] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf. *Communications of the ACM*, 65:99 – 106, 2020. 2, 5
- [40] Jesús R Nieto and Antonio Susín. Cage based deformations: a survey. In *Deformation Models: Tracking, Animation and Applications*, pages 75–99. Springer, 2012. 2, 3
- [41] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. *ArXiv*, abs/2312.05941, 2023. 3
- [42] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 5
- [43] Keunhong Park, U. Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5845–5854, 2020. 2
- [44] Keunhong Park, U. Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf. *ACM Transactions on Graphics (TOG)*, 40:1 – 12, 2021. 2
- [45] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9050–9059, 2020. 2
- [46] Yicong Peng, Yichao Yan, Shengqi Liu, Yuhao Cheng, Shanyan Guan, Bowen Pan, Guangtao Zhai, and Xiaokang Yang. Cagenerf: Cage-based neural radiance field for generalized 3d deformation and animation. In *NeurIPS*, 2022. 3
- [47] Malte Prinzler, Otmar Hilliges, and Justus Thies. Diner: Depth-aware image-based neural radiance fields. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12449–12459, 2022. 2
- [48] Sergey Prokudin, Qianli Ma, Maxime Raafat, Julien Valentin, and Siyu Tang. Dynamic point fields. *arXiv preprint arXiv:2304.02626*, 2023. 2
- [49] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians, 2023. 2, 3

- [50] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. 2024. 6, 7, 8
- [51] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001. 3
- [52] Edoardo Remelli, Timur M. Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabián Prada, Jason M. Saragih, and Yaser Sheikh. Drivable volumetric avatars using texel-aligned features. *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 2, 6
- [53] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. *ArXiv*, abs/2312.03704, 2023. 3
- [54] Hang Si. Tetgen: A quality tetrahedral mesh generator and a 3d delaunay triangulator (version 1.5 — user’s manual). 2013. 4
- [55] Shih-Yang Su, Frank Yu, Michael Zollhoefer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Neural Information Processing Systems*, 2021. 2
- [56] Shih-Yang Su, Timur M. Bagautdinov, and Helge Rhodin. Danbo: Disentangled articulated neural body representations via graph neural networks. In *European Conference on Computer Vision*, 2022.
- [57] Shih-Yang Su, Timur M. Bagautdinov, and Helge Rhodin. Npc: Neural point characters from video. *ArXiv*, abs/2304.02013, 2023. 2
- [58] Robert W. Sumner and Jovan Popović. Deformation transfer for triangle meshes. *ACM SIGGRAPH 2004 Papers*, 2004. 3, 4
- [59] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 6105–6114. PMLR, 2019. 3
- [60] Ayush Tewari, Otto Fried, Justus Thies, Vincent Sitzmann, S. Lombardi, Z. Xu, Tanaba Simon, Matthias Nießner, Edgar Tretschk, L. Liu, Ben Mildenhall, Pranatharthi Srinivasan, R. Pandey, Sergio Orts-Escolano, S. Fanello, M. Guang Guo, Gordon Wetzstein, J y Zhu, Christian Theobalt, Manju Agrawala, Donald B. Goldman, and Michael Zollhöfer. Advances in neural rendering. *Computer Graphics Forum*, 41, 2021. 2
- [61] Ayush Kumar Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason M. Saragih, Matthias Nießner, Rohit Pandey, S. Fanello, Gordon Wetzstein, Jun-Yan Zhu, Christian Theobalt, Maneesh Agrawala, Eli Shechtman, Dan B. Goldman, and Michael Zollhofer. State of the art on neural rendering. *Computer Graphics Forum*, 39, 2020. 2
- [62] Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. Styleavatar: Real-time photo-realistic portrait avatar from a single video. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 7
- [63] Yifan Wang, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Transactions on Graphics (TOG)*, 38:1 – 14, 2019. 2, 3
- [64] Yifan Wang, Noam Aigerman, Vladimir G. Kim, Siddhartha Chaudhuri, and Olga Sorkine-Hornung. Neural cages for detail-preserving 3d deformations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 72–80. Computer Vision Foundation / IEEE, 2020. 3
- [65] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humanerf: Free-viewpoint rendering of moving people from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16189–16199. IEEE, 2022. 2
- [66] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Wang Xinggang. 4d gaussian splatting for real-time dynamic scene rendering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3
- [67] Donglai Xiang, Fabián Prada, Timur M. Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica K. Hodgins, and Chenglei Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM Transactions on Graphics (TOG)*, 40:1 – 15, 2021. 6
- [68] Donglai Xiang, Fabián Prada, Zhe Cao, Kaiwen Guo, Chenglei Wu, Jessica K. Hodgins, and Timur M. Bagautdinov. Drivable avatar clothing: Faithful full-body telepresence with dynamic clothing driven by sparse rgb-d input. 2023. 2
- [69] Jun Xiang, Xuan Gao, Yudong Guo, and Ju yong Zhang. Flashavatar: High-fidelity digital avatar rendering at 300fps. *ArXiv*, abs/2312.02214, 2023. 2, 3
- [70] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics, 2023. 3
- [71] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5428–5438, 2022. 2
- [72] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. *ArXiv*, abs/2312.03029, 2023. 2, 3
- [73] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2024. 3
- [74] Hao Zhang, Yao Feng, Peter Kulits, Yandong Wen, Justus Thies, and Michael J. Black. Text-guided generation and editing of compositional 3d avatars. *ArXiv*, abs/2309.07125, 2023. 2

- [75] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6
- [76] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. *ArXiv*, abs/2312.02155, 2023. 2, 3
- [77] Yufeng Zheng, Yifan Wang, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21057–21067, 2022. 2
- [78] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15872–15882, 2022. 2
- [79] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4574–4584, 2022. 2
- [80] Wojciech Zielonka, Timo Bolkart, Thabo Beeler, and Justus Thies. Gaussian eigen models for human heads. *arXiv:2407.04545*, 2024. 3
- [81] Wojciech Zielonka, Stephan J. Garbin, Alexandros Lattas, George Kopanas, Paulo Gotardo, Thabo Beeler, Justus Thies, and Timo Bolkart. Synthetic prior for few-shot drivable head avatar inversion. *arXiv:2501.06903*, 2025. 3
- [82] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. *Computer Graphics Forum*, 37, 2018. 2
- [83] Matthias Zwicker, Hans Rüdiger Pfister, Jeroen van Baar, and Markus H. Gross. Surface splatting. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001. 3

## A.4 GAUSSIAN EIGEN MODELS FOR HUMAN HEADS

*Gaussian Eigen Models for Human Heads*

Wojciech Zielonka, Timo Bolkart, Thabo Beeler, Justus Thies

Published in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA, 2025.

## Abstract

Current personalized neural head avatars face a trade-off: lightweight models lack detail and realism, while high-quality, animatable avatars require significant computational resources, making them unsuitable for commodity devices. To address this gap, we introduce Gaussian Eigen Models (GEM), which provide high-quality, lightweight, and easily controllable head avatars. GEM utilizes 3D Gaussian primitives for representing the appearance, combined with Gaussian splatting for rendering. Building on the success of mesh-based 3D morphable face models (3DMM), we define GEM as an ensemble of linear eigenbases for representing the head appearance of a specific subject. In particular, we construct linear bases to represent the position, scale, rotation, and opacity of the 3D Gaussians. This allows us to efficiently generate Gaussian primitives of a specific head shape by a linear combination of the basis vectors, only requiring a low-dimensional parameter vector that contains the respective coefficients. We propose to construct these linear bases (GEM) by distilling high-quality, compute-intensive CNN-based Gaussian avatar models that can generate expression-dependent appearance changes like wrinkles. These high-quality models are trained on multi-view videos of a subject and are distilled using a series of principal component analyses. Once we have obtained the bases that represent the animatable appearance space of a specific human, we learn a regressor that takes a single RGB image as input and predicts the low-dimensional parameter vector that corresponds to the shown facial expression. In a series of experiments, we compare GEM’s self-reenactment and cross-person reenactment results to state-of-the-art 3D avatar methods, demonstrating GEM’s higher visual quality and better generalization to new expressions.



# Gaussian Eigen Models for Human Heads

Wojciech Zielonka<sup>1,2</sup> Timo Bolkart<sup>3</sup> Thabo Beeler<sup>3</sup> Justus Thies<sup>1,2</sup>

<sup>1</sup>Max Planck Institute for Intelligent Systems, Tübingen, Germany

<sup>2</sup>Technical University of Darmstadt <sup>3</sup>Google

<https://zielon.github.io/gem/>

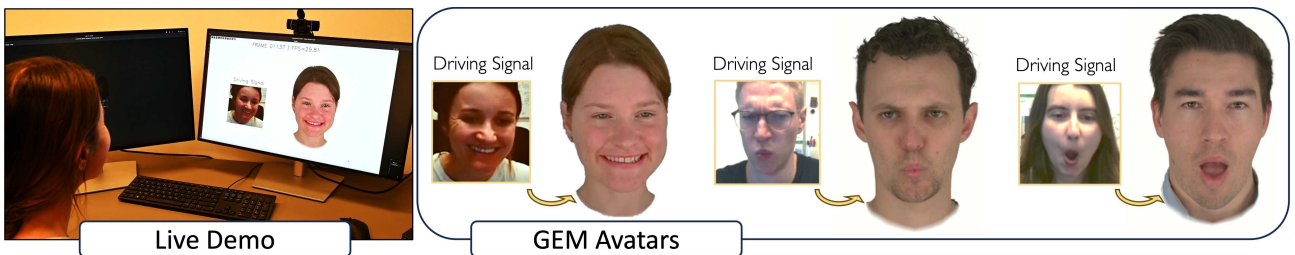


Figure 1. We propose a method that represents 3D Gaussian head avatars in a network-free form as ensembles of eigenbases (GEM). Only a linear combination of these bases is needed to generate new primitives, which can be splatted using 3D Gaussian Splatting. We demonstrate that the necessary coefficients for a specific expression can be regressed from single images, enabling real-time facial animation and cross-reenactment. The simplicity of GEM results in highly efficient storage and rendering times.

## Abstract

Current personalized neural head avatars face a trade-off: lightweight models lack detail and realism, while high-quality, animatable avatars require significant computational resources, making them unsuitable for commodity devices. To address this gap, we introduce Gaussian Eigen Models (GEM), which provide high-quality, lightweight, and easily controllable head avatars. GEM utilizes 3D Gaussian primitives for representing the appearance combined with Gaussian splatting for rendering. Building on the success of mesh-based 3D morphable face models (3DMM), we define GEM as an ensemble of linear eigenbases for representing the head appearance of a specific subject. In particular, we construct linear bases to represent the position, scale, rotation, and opacity of the 3D Gaussians. This allows us to efficiently generate Gaussian primitives of a specific head shape by a linear combination of the basis vectors, only requiring a low-dimensional parameter vector that contains the respective coefficients. We propose to construct these linear bases (GEM) by distilling high-quality compute-intense CNN-based Gaussian avatar models that can generate expression-dependent appearance changes like wrinkles. These high-quality models are trained on multi-view videos of a subject and are dis-

tilled using a series of principle component analyses.

Once we have obtained the bases that represent the animatable appearance space of a specific human, we learn a regressor that takes a single RGB image as input and predicts the low-dimensional parameter vector that corresponds to the shown facial expression. We demonstrate that this regressor can be trained such that it effectively supports self- and cross-person reenactment from monocular videos without requiring prior mesh-based tracking. In a series of experiments, we compare GEM’s self-reenactment and cross-person reenactment results to state-of-the-art 3D avatar methods, demonstrating GEM’s higher visual quality and better generalization to new expressions. As our distilled linear model is highly efficient in generating novel animation states, we also show a real-time demo of GEMs driven by monocular webcam videos. The code and model will be released for research purposes.

## 1. Introduction

Half a century ago, Frederick Parke described a representation and animation technique to generate „animated sequences of a human face changing expressions” [40]. Using polygonal meshes, single facial expression states were described that could be combined with linear interpolation

to generate new expression states (the „simplest way consistent with natural motion” [40]). Based on this principle, Blanz, and Vetter [2] introduced the so-called 3D morphable model (3DMM) - a statistical model of the 3D shape and appearance of human faces. Principle Component Analysis (PCA) is performed on a set of around 200 subjects that have been laser-scanned and registered to a consistent template to find the displacement vectors (principal components) of how faces change the most, in terms of geometry and albedo. With this PCA basis, new faces can be generated by specifying the coefficients for the principle components taking a dot product of the coefficients with the basis to obtain offsets, and adding them to the mean. State-of-the-art reports on face reconstruction and tracking [68] as well as on morphable models [6] state that this representation is widely used for facial performance capturing (regression-based and optimization-based) and builds the backbone of recent controllable photo-realistic 3D avatars that are equipped with neural rendering [10, 15, 48, 49, 62].

Inspired by the simplicity of such mesh-based linear morphable models and addressing the lack of appearance realism of current 3DMMs, we propose a personalized linear appearance model based on 3D Gaussians as geometry primitives following 3D Gaussian Splatting (3DGS) [21]. In contrast to the work on Dynamic 3D Gaussian Avatars [30, 38, 43, 45, 57, 60, 66], our goal is a compact and light representation that does not need vast amounts of compute resources to generate novel expressions of the human. Unfortunately, most of the methods show that to produce high-quality results, one needs to employ heavy CNN-based architectures which are not well suited for commodity devices and tend to slow down the rendering pipeline. Moreover, those models comprise dozens of millions of parameters creating heavy checkpoints that can easily exceed 500 MB. This ultimately creates a major issue for distributing and managing personalized models. We tackle this problem by distilling a CNN-based architecture, leading to a personalized **Gaussian Eigen Models for Human Heads, GEM** in short. Our approach builds on Gaussian maps predicted from a modified UNet architecture [53] which is used for the UV space normalization required to build linear eigenbases. Based on the per-subject trained CNN model, we bootstrap the GEM by computing an ensemble of linear bases on the predicted Gaussian maps of the training frames. The bases are refined on the training corpus using photometric losses while preserving their orthogonality.

These lightweight appearance bases are controlled with a relatively low number of parameters ranging from twenty up to fifty coefficients which can be specified w.r.t. the available compute resources and can for example be regressed by a ResNet-based model [8]. We demonstrate this for self-reenactment as well as cross-person animation, including a real-time demo in the suppl. video.

In summary, our main contributions are:

1. Gaussian Eigen Models for Human Heads (GEM), a distillation technique of 3D Gaussian head avatar models built upon an ensemble of eigenbases.
2. real-time (cross-person) animation of GEMs from single input images using a generalizable regressor.

## 2. Related Work

The majority of face representation and tracking techniques are based on parametric 3D morphable models (3DMM) [2, 29]. For a detailed overview, we refer to the state-of-the-art reports on face tracking and reconstruction [68], the report on morphable models [6], and the two neural rendering state-of-the-art reports [48, 49] that demonstrate how neural rendering can be leveraged for photo-realistic facial or full body avatars. Next, we review the recent methods for photo-realistic 3D avatars generation which build appearance models using neural radiance fields (NeRF) [35] or volumetric primitives like 3D Gaussians [21].

### 2.1. NeRF-based avatars

One of the first methods that combines a 3DMM and NeRF is NeRFace [10], where a neural radiance field is directly conditioned by expression codes of the Basel Face Model (BFM) [2, 50]. This idea gave rise to many methods [11, 15, 42, 56, 59, 62–64] following a similar approach, but attaching the radiance fields more explicitly to the surface of the 3DMM, e.g., by using the 3DMM-defined deformation field. For photorealistic results, some methods employ StyleGAN2-like architectures [20] with a NeRF-based renderer [1, 4, 19]. Generative methods like EG3D [4] and PanoHead [1] employ GAN-based training to predict tri-plane features that span a NeRF. GANAvatar [19] applies this scheme to reconstruct a personalized avatar.

Close to our method is StyleAvatar [53]. Based on 3DMM tracking the method learns a personalized avatar that benefits from a StyleUNet which incorporates StyleGAN [20] to decode the final image. Despite real-time capabilities, StyleAvatar suffers from artifacts produced by the image-to-image translation network that we explicitly avoid by using Gaussian maps which can compensate for tracking misalignments by predicting corrective fields for the 3D Gaussians.

### 2.2. 3D Avatars from Volumetric Primitives

Using multiview images with a variational auto-encoder [22] and volumetric integration, Neural Volumes (NV) [31] encodes dynamic scenes into a volume which can be deformed by traversing a latent code  $\mathbf{z}$ . To better control the 3D space, Lombardi *et al.* [32] introduce Mixture of Volumetric Primitives (MVP) a hybrid representation based on primitives attached to a tracked mesh which ultimately replaced the encoder from NV. Each primitive is a volume

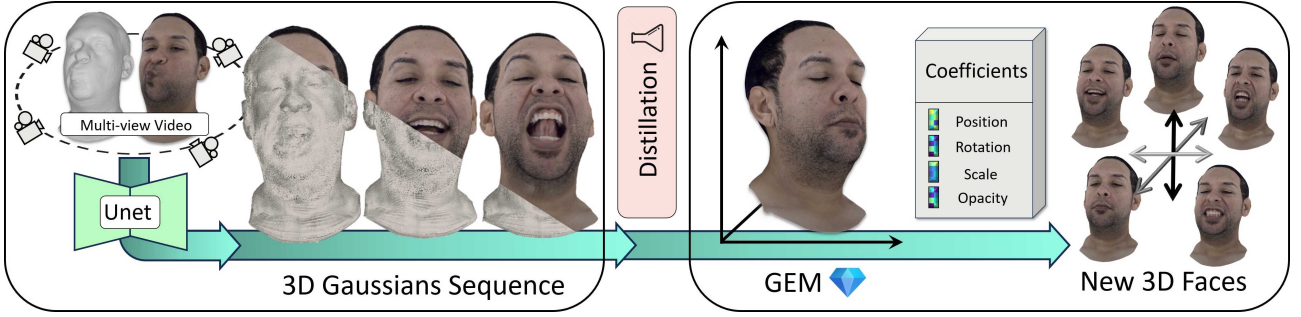


Figure 2. Given a multi-view video of a subject and mesh tracking, we create a dataset of 3D Gaussian point clouds for each frame in the sequence. Using this data, we distill a high-quality Gaussian Eigen Model (GEM). GEM is an ensemble of linear bases for each Gaussian primitive modality: position, opacity, scale, and rotation. Based on these bases, facial appearances are generated by a linear combination.

represented as a small voxel with  $32^3$  cells that store RGB and opacity values. The final color is obtained by integrating values along a pixel ray. This hybrid representation inspired many follow-up projects [3, 28, 44, 47, 54]. As an alternative to MVP primitives, 3D Gaussian Splatting (3DGS) [21] represents a volume as a set of anisotropic 3D Gaussians, which are equivalently described as ellipsoids, in contrast to isotropic spheres used in Pulsar [25].

Numerous methods [9, 13, 17, 24, 30, 38, 43, 45, 55, 57, 61, 66, 67] capitalize on the speed and quality of 3DGS. Qian *et al.* [43] attach 3D Gaussians to the FLAME [29] mesh surface and apply a deformation gradient similar to Zielenka *et al.* [64] to orient the Gaussians according to the local Frenet frames of the surface. This method, however, does not utilize any information about expressions and, thus, struggles with pose-dependent changes (e.g., wrinkles, self-shadows) and, despite high-quality results, retrieves only a global static appearance model. 3D Gaussian blendshapes [34] controls an avatar by linearly interpolating between optimized blendshapes using 3DMM expression coefficients. However, this method depends on an underlying 3DMM whereas GEM is a mesh-free representation. Li *et al.* [30] use a StyleUNet-like CNN architecture [53] to regress front and back Gaussian maps. Employing a powerful CNN network on position maps, they achieve impressive results for human bodies with effects like pose-dependent wrinkle formation.

Please note that in this work, we focus on methods that directly output Gaussian primitives. This is an important distinction from a branch of methods that follow Deferred Neural Rendering [51], where a refinement CNN translates splatted features or coarse colors into the final image; for instance, Gaussian Head Avatars [57] and NGPA [13]. This distinction is important because Gaussian primitives cannot be fully distilled into an eigenbasis in this context, as the refinement CNN network is required to complete the rendering directly in the image space.

### 2.3. 3DGS Compression Methods

Recently, several methods [7, 14, 26, 27, 36, 39] have been proposed to reduce the memory footprint of 3D Gaussian Splatting (3DGS). Papantonakis *et al.* [39] apply codebook quantization to the Gaussian primitive properties, alongside pruning of Spherical Harmonic (SH) coefficients based on their final contribution. In contrast to postprocessing approaches [7, 27, 39], Compact3D [36] employs a single-stage process that jointly optimizes both the codebook entries and the primitives. Fan *et al.* [7] calculate a significance score for each primitive by measuring its pixel hit count, thereby improving the pruning strategy. Most of these methods target static scenes or time-conditioned environments, unlike our approach, which focuses on efficient, fully controllable head avatars. Nonetheless, these compression techniques could be adapted to our animatable avatars to reduce memory usage.

### 3. Method

Recent dynamic 3D Gaussian Avatar methods show unprecedented quality, however, they require sophisticated and often compute-heavy CNN-based architectures [30, 38, 57] to capture high-frequency and dynamic details like pose-dependent wrinkles or self-shadows. The aim of this paper is to build on top of this quality but remove the compute-intense architecture during inference. Specifically, we propose to distill high-quality avatar models into lightweight linear animation models which we call GEMs. A GEM is defined by an ensemble of eigenbases that span the space of the 3D Gaussian primitives. These eigenbases are constructed via PCA applied on a dataset of per-frame Gaussian primitives, see Section 3.1.

*An important distinction compared to other neural avatars [10, 15, 31, 53, 57, 64] is that GEM does not require a 3DMM [29, 41] at test time.* We demonstrate that a GEM can be directly driven by a monocular video using a generalized image-based regression network, see Section 3.2.



### 3.1. Gaussian Eigen Model (GEM)

For our distillation, we reconstruct a sequence of normalized Gaussian primitives  $\mathcal{D} = \{\mathbf{G}_0, \dots, \mathbf{G}_{N-1}\}$ . As input, we assume a multi-view video of the subject with  $N$  time frames. Per time frame  $i$ , we reconstruct the 3D Gaussian pointcloud  $\mathbf{G}_i$ , where  $\mathbf{G}_i$  contains the parameters that define the 3D Gaussians such as rotation  $\theta$ , position  $\phi$ , opacity  $\alpha$ , scale  $\sigma$ , and color  $\vec{c}$  such that  $\mathbf{G}_i = \{\vec{\theta}, \vec{\phi}, \vec{\alpha}, \vec{\sigma}, \vec{c}\}$ .

#### Reconstructing High-quality 3D Gaussian Primitives:

We are following the idea of organizing the 3D Gaussians in 2D maps [30, 38, 45, 57], where each pixel represents a 3D Gaussian with its parameters. We propose an adapted CNN-architecture of Animatable Gaussians (AG) [30], by merging the separate Style-U-Nets, reducing the convolutional layers, and operating in the UV space of the FLAME head model. In addition, we are employing deformation gradients following Sumner *et al.* [46] to handle the transformation from canonical to deformed space and treat the color as a global parameter. We refer to the suppl. mat. for a detailed explanation of the architectural changes. In comparison to the original AG model, our proposed CNN model produces slightly better results while being more efficient in terms of computing and memory. Using this model, we generate the per-frame Gaussian primitives  $\mathbf{G}_i$  in the canonical space for all training time-frames. Note that for this reconstruction, we follow Animatable Gaussians and, thus, FLAME-based tracking is required. However, during inference, our model is independent of FLAME.

**Distillation:** Given  $\mathcal{D} = \{\mathbf{G}_0, \dots, \mathbf{G}_{N-1}\}$ , we build a personalized eigenbasis model, which is called GEM. We compute a statistical model for each Gaussian modality separately. Specifically, we create individual bases for rotation  $\mathbf{B}_\theta$ , position  $\mathbf{B}_\phi$ , opacity  $\mathbf{B}_\alpha$ , and scale  $\mathbf{B}_\sigma$  with respective means  $\vec{\mu}_\theta$ ,  $\vec{\mu}_\phi$ ,  $\vec{\mu}_\alpha$  and  $\vec{\mu}_\sigma$  via Principle Component Analysis (PCA) [18]. Note that the color  $\mathbf{C}$  is optimized globally and, thus, acts as a classical texture without the need to apply PCA. To accurately learn dynamically moving Gaussians, we fixed the color to prevent it from dominating the image representation, otherwise, Gaussians could change their semantic meaning (e.g., a Gaussian could represent the lip in one state, and the teeth in the other deformation state). Keeping the semantic meaning of specific Gaussians across deformation states is crucial for applying a PCA afterward.

A face model instance  $\mathbf{G}$  is represented as a linear combination of these bases:

$$\mathbf{G} = \{\vec{\mu}_i + \mathbf{B}_i \mathbf{k}_i \mid i \in \{\theta, \phi, \alpha, \sigma\}, \vec{c}\}, \quad (1)$$

where  $\mathbf{k}_\theta$ ,  $\mathbf{k}_\phi$ ,  $\mathbf{k}_\sigma$  and  $\mathbf{k}_\sigma \in \mathbb{R}^M$  are the linear coefficients which are defining the facial expression state, assuming  $M$  principal components. As an example, Figure 3 shows posi-



Figure 3. **Samples of a GEM.** We display samples for the first three components of the position  $\mathbf{k}_\phi$  eigenbasis of a GEM, showing diverse expressions. Note that GEM requires **no** parametric 3D face model like FLAME[29].

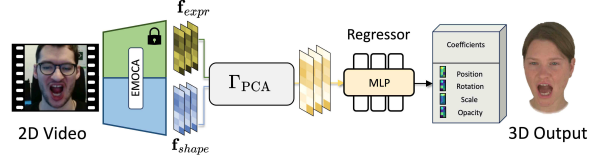


Figure 4. **Image-based animation.** One of the applications of our GEM is real-time (cross)-reenactment. For that, we utilize generalized features from EMOCA [5] and build a pipeline to regress the PCA coefficients of our model from an input image/video.

tion parameter  $\mathbf{k}_\phi$  sampled in the range of  $[-3\sigma_\phi, 3\sigma_\phi]$  ( $\sigma_\phi$  being the std. deviation).

As the Gaussian primitives  $\mathcal{D}$  might contain tracking failures and misalignments, the principle components  $\mathbf{B}_{(\theta, \phi, \alpha, \sigma)}$  also contain artifacts as well. We, therefore, refine the bases using the training images directly, by applying a photometric reconstruction loss. We employ the same objectives from the CNN model training (see suppl. mat.).

$$\mathcal{L}_{Color} = (1 - \omega)\mathcal{L}_1 + \omega\mathcal{L}_{D-SSIM} + \zeta\mathcal{L}_{VGG} \quad (2)$$

We refine the base vectors for around 30k iterations. To ensure that the individual bases stay orthonormal throughout this refinement, every 1k steps, we orthogonalize the bases using QR decomposition. This refinement improves the training PSNR errors from 34.75dB to 36.68dB and 36.85dB for the training steps 0k, 5k, and 30k, respectively. Throughout our experiments, we did not encounter overfitting issues with this scheme. The reconstruction metrics on two randomly selected test sequences with refinement are: PSNR: **31.51**, LPIPS: **0.091**, SSIM: **0.936**; and without: PSNR: 31.38, LPIPS: 0.094, SSIM: 0.933.

### 3.2. Image-based Animation

Expressions for a GEM are fully defined by their coefficients  $\mathbf{k}_\theta$ ,  $\mathbf{k}_\phi$ ,  $\mathbf{k}_\sigma$  and  $\mathbf{k}_\sigma$ . This is a similar idea to codec avatars [33], however, our approach does not need additional pixel shaders in the form of a small regression MLP. There are several ways to obtain the coefficients of a GEM, for example, one can employ analysis-by-synthesis-based optimization or regression. Analysis-by-synthesis [2] is the backbone of current avatar methods, as they use photomet-

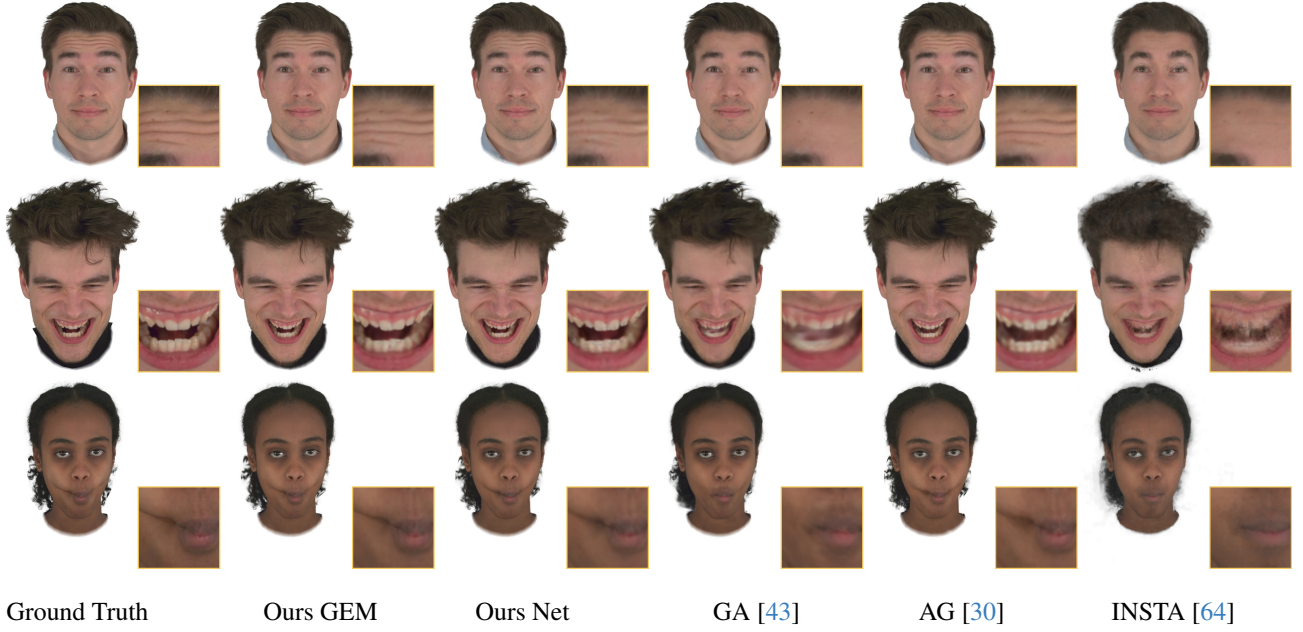


Figure 5. **Novel view synthesis.** Both, our CNN and GEM show better performance on novel views, especially, in the region of the mouth interior and wrinkles. In this experiment, we are following the evaluation of Gaussian Avatars [43] and demonstrate novel viewpoint generation. GEM is obtained throughout analysis-by-synthesis fitting [2, 50]. Note that the expressions are seen during training.

ric or depth-based face trackers to sequentially optimize the coefficients of the underlying 3DMMs like FLAME [15, 50, 64] which is typically slow. As a fast, but more imprecise alternative, regressors like DECA [8] or EMOCA [5] can be used which are built on a ResNet backbone and regress FLAME parameters directly from an image. We apply several modifications to the EMOCA model, see 4. We use intermediate features of the pre-trained EMOCA network denoted as  $\Theta(\mathbf{I}_i)$  where  $\mathbf{I}_i$  is the current image. EMOCA’s architecture comprises two ResNet networks; one to extract expression features  $\mathbf{f}_{expr} \in \mathbb{R}^{2048}$  and the second for shape  $\mathbf{f}_{shape} \in \mathbb{R}^{2048}$ , both are followed by final MLPs to regress corresponding FLAME parameters. As we do not rely on FLAME, we remove the last hidden layer of the final MLP obtaining two feature vectors which we combine into one  $\mathbf{f} \in \mathbb{R}^{2 \times 1024}$  vector. For these features, we build a PCA layer with a basis denoted as  $\hat{\mathbf{R}}$  using the training frames from five frontal cameras of NeRSemble. Note that we use relative features  $\mathbf{r} = \mathbf{f} - \mathbf{f}_{neutral}$  in this PCA layer. The neutral reference frame  $\mathbf{f}_{neutral} = \Theta(\mathbf{I}_{neutral})$  to compute these relative features is selected manually from the video, similar to Face2Face [50]. During training, for each frame, we project  $\mathbf{r}$  onto the PCA manifold using the first 50 principal components to restrict and regularize training. Finally, we use their corresponding PCA coefficients:

$$\kappa = (\mathbf{r} - \bar{\mathbf{R}})\hat{\mathbf{R}}^T, \quad (3)$$

where  $\bar{\mathbf{R}}$  is the relative PCA model mean. The projected coefficients are passed through a small MLP that produces a vector of GEM coefficients  $\mathbf{k} = \{\mathbf{k}_\theta, \mathbf{k}_\phi, \mathbf{k}_\sigma\}$ :

$$\mathbf{k} = 3 \cdot \sigma_k \cdot \tanh(\text{MLP}(\kappa)). \quad (4)$$

The MLP has three hidden layers with 256 neurons each and ReLU activations. We use a scaled tanh activation function for the output to restrict the prediction to be in  $[-3 \cdot \sigma_k, 3 \cdot \sigma_k]$ ,  $\sigma_k$  being the respective standard deviation of the coefficients  $\mathbf{k}$ , obtained from the PCA. The final primitives are obtained by Eq. 1 and splatted using 3DGS.

## 4. Results

We evaluate GEM on the NeRSemble [23], where tracked meshes [43] and synchronized images from 16 cameras with a resolution of  $802 \times 550$  are available. Our baselines are Gaussian Avatars (GA) [43] which is neural network-free (Gaussians are attached to the FLAME model), our implementation of Animatable Gaussians (AG) [30] which is based on CNN-predicting Gaussian maps, and INSTA [64] which uses dynamic NeRF [35]. Note that all baselines require at least two stages: (i) construct the avatar, and (ii) get the parameters to drive it. Most of them use offline tracking with additional objectives like hair reconstruction [12, 43], which does not work for real-time applications despite the avatar model’s rendering being real-time. Importantly, in our approach, we introduce a third step, i.e., the construction of the eigenbasis (GEM), which only introduces **negligible** computational costs ( $\sim 1$  min) in comparison to the avatar reconstruction itself. For the comparison, we present both of our appearance models, the StyleUNet-based architecture (**Ours Net**) and the distilled linear Eigen model (**Ours GEM**) which we evaluate using analysis-by-synthesis fitting to the target images following [2, 50, 65]. Additionally, we present cross-reenactment results based on our coefficient regressor, compared to the baselines that

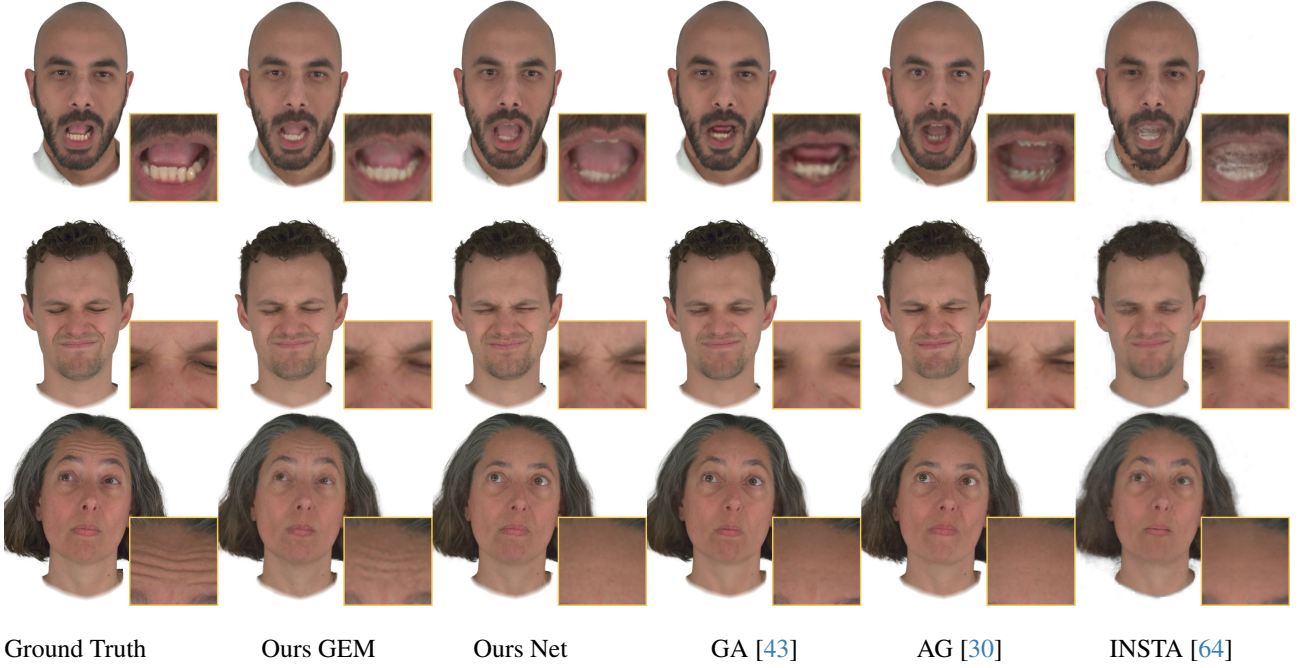


Figure 6. **Novel view and expression synthesis.** Our Gaussian Eigen Models for Human Heads shows better results in regions like teeth, wrinkles, and self-shadows compared to other methods that struggle with artifacts.

Method	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	L1 $\downarrow$
AG [30]	32.4166	0.0712	0.9614	0.0066
GA [43]	31.3197	0.0786	0.9567	0.0075
INSTA [64]	27.7786	0.1232	0.9294	0.0163
Ours Net	32.4622	0.0713	0.9617	0.0067
Ours GEM	<b>33.5528</b>	<b>0.0678</b>	<b>0.9662</b>	<b>0.0061</b>

Table 1. **Novel viewpoint evaluation** is conducted on a withhold camera from the 16 cameras used for training. Note that the expression has been seen during training, and only the view is new.

Method	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	L1 $\downarrow$
AG [30]	29.0114	0.0812	0.9429	0.0099
GA [43]	28.3137	0.0815	0.9433	0.0102
INSTA [64]	27.9181	0.1153	0.9340	0.0128
Ours Net	29.2454	0.0777	0.9448	0.0096
Ours GEM	<b>32.6781</b>	<b>0.0675</b>	<b>0.9633</b>	<b>0.0069</b>

Table 2. **Evaluation on novel expressions** and views show improved results of GEM optimized using analysis-by-synthesis compared to others. Figure 6 shows the corr. qualitative results.

use FLAME meshes regressed by DECA [8]. Relative expression transfer based on ground truth meshes [43] can be found in the supp. mat. All of the methods are evaluated using several image space metrics on novel expressions and novel views, following the test and novel-view split of Qian *et al.* [43]. For our GEM models, we use 50 components distilled from  $256^2$  textures which give around 60k active Gaussians. Animatable Gaussians [30] uses a similar amount of primitives for front and back textures and Gaussian Avatars [43] around 100k Gaussians.

#### 4.1. Image Quality Evaluation

To evaluate our method, we measure the color error in the image space using the following metrics: PSNR (dB), LPIPS [58], L1 loss, and structural similarity (SSIM). We follow the evaluation scheme from Gaussian Avatars [43], using their train and validation split. The evaluation of GEM was generated by sequentially fitting the coefficients to each image using photometric objectives. Note that the baselines use the FLAME model with offsets for the track-

ing, while GEM can directly be used for tracking.

Table 2 presents results on novel expressions evaluated on all 16 cameras. Both the quantitative and qualitative results depicted in Figure 6 show that our PCA model produces fewer artifacts, especially for regions like teeth or facial wrinkles. Table 1 contains an evaluation where we measure errors on novel viewpoints. The results demonstrate that our CNN-based appearance model outperforms other neural methods, while our linear eigenbasis GEM achieves the highest quality. This is due to the 'direct' analysis-by-synthesis approach, which fully leverages the expressiveness and detail of our photorealistic appearance model, without the limitations imposed by 3DMMs such as FLAME. Moreover, Figure 5 shows qualitative results of our method on novel views. As can be seen, we better capture high-frequency details, pose-dependent wrinkles, and self-shadows - something which is not possible for methods like Gaussian Avatars [43] or INSTA [64], since they either do not use expression-dependent neural networks or limit the conditioning to a small region only.



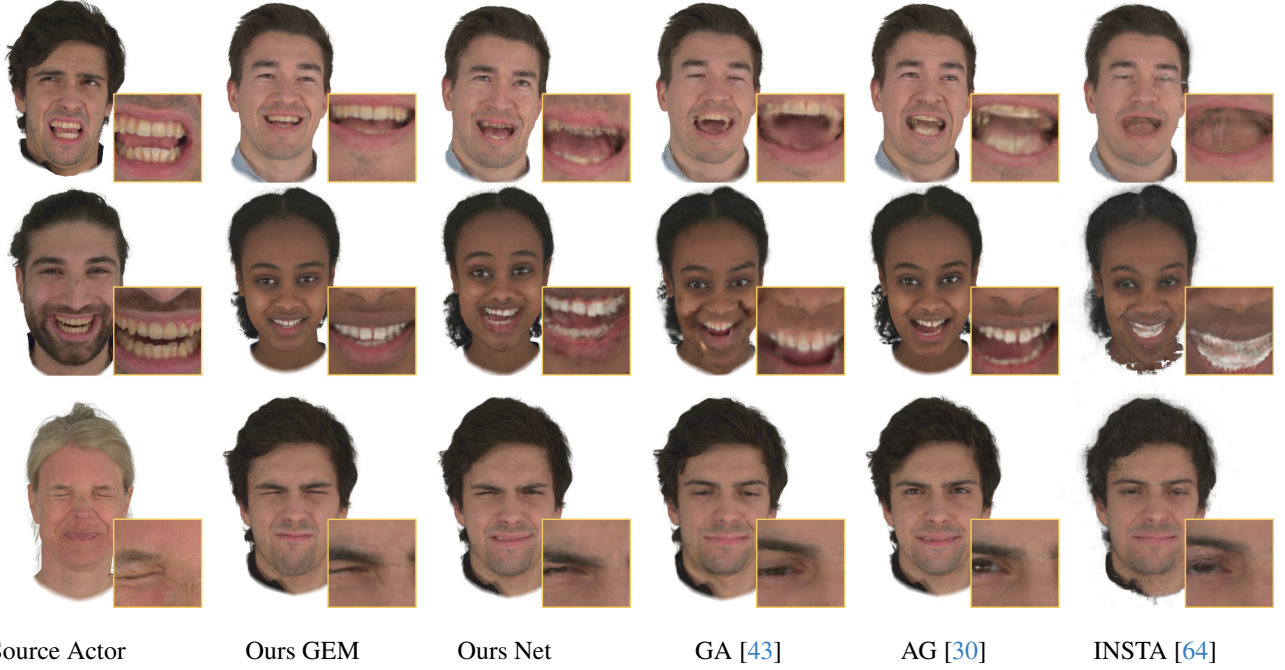


Figure 7. **Facial cross-person reenactment using an image-based regressor.** The reenactment of the baselines is performed using relative transfer between FLAME meshes regressed by EMOCA compared to our GEM regressor network (Ours GEM).

## 4.2. Cross-person Reenactment Evaluation

Facial cross-person reenactment transfers expressions from the source actor to the target actor. For this, the baseline methods require tracked meshes obtained by fitting the 3DMM model for each frame of the source actor sequence. As an alternative to optimization-based tracking, a (monocular) regressor like EMOCA [5], can predict such tracked meshes in real-time. We demonstrate this in Figure 7, where GEM is driven by our image-based regressor and the others by EMOCA. As shown, our network-based method and GEM produce sharp results, while the baseline methods struggle to extrapolate to new expressions, displaying severe artifacts in appearance. Our approach effectively regularizes the regressed coefficients, ensuring that the predicted avatar remains in the training distribution and thereby avoids artifacts seen in INSTA or Gaussian Avatars. Drawing inspiration from EMOCA [5], we further assess cross-re-enactment quantitatively by leveraging emotion recognition feature vectors from both the source image and the resulting cross-re-enactment, utilizing EmoNet [52]. For each

Method	$E_{feat}^{cos} \uparrow$	$E_{feat}^{L_1} \downarrow$	FID $\downarrow$	FPS $\uparrow$
AG	0.9396	5.3399	0.4093	16.51
GA	0.8917	6.6141	0.5593	142.71
INSTA	0.9087	6.3153	0.5299	20.62
Ours Net	0.9440	5.1044	0.3685	35.77
Ours GEM	0.9381	5.3197	0.4286	201.70

Table 3. **Cross-reenactment evaluation** employing EmoNet features and FID score.

pair of input and output images, we predict EmoNet features and measure cosine distance and  $L_1$  error between them. We report the numbers in the Table 3. Additionally, we also report FID scores [16] and rendering speed. Our method achieves on-par quality with the CNN-based solution while maintaining the highest frame rates and outperforming GA in terms of quality.

## 4.3. GEM Ablation Studies

We are interested in the compression error introduced by the projection on different amounts of principal components used in GEM, also concerning the memory consumption. Our smallest model weighs as little as **7MB** using only 10 components of the eigenbasis. This is almost **12** times less than our smallest CNN-based model and almost **70** times less than Animatable Gaussians [30]. In contrast to neural networks, we can easily trade quality over size which is very useful in the context of different commodity devices with reduced compute capabilities. Table 4 presents how compression affects the quality of reconstruction, where we evaluate a sequence with  $\sim 1k$  frames for a single actor under a novel view. As expected, using only 10 components impacts the quality the most, however, the results are still of high quality, see Figure 8. Gaussian Avatars [43] offers a small size of the stored Gaussians cloud, ranging from 5MB, and 14MB without the FLAME model for  $128^2$ ,  $256^2$  Gaussians, respectively. However, the quality of reconstruction lacks wrinkle details and sharpness as can be seen in Figure 10. In comparison to Gaussian Avatars [43], our model does not require FLAME during inference which is an additional **90MB**.

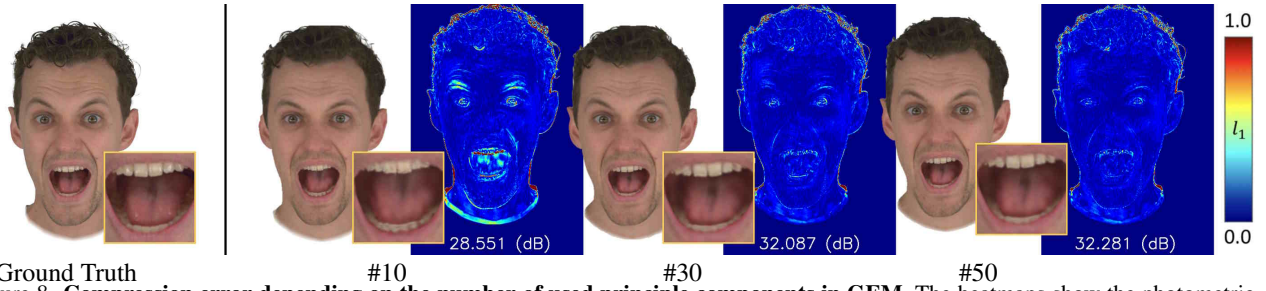


Figure 8. **Compression error depending on the number of used principle components in GEM.** The heatmaps show the photometric  $\ell_1$ -error for 10, 30, and 50 components using  $128^2$  Gaussian maps. See suppl. doc. for additional evaluations.



Figure 9. Despite fixed topology and predefined texture size GEM faithfully represents facial attributes like glasses.



Figure 10. The quality comparison to Gaussian Avatars. Note that we do **not** need an additional FLAME model which weighs 90MB.

#Comp	128 <sup>2</sup>			256 <sup>2</sup>			512 <sup>2</sup>		
	PSNR ↑	Size MB	FPS ↑	PSNR ↑	Size MB	FPS ↑	PSNR ↑	Size MB	FPS ↑
10	31.81	7	237.96	31.88	28	210.03	32.23	113	130.46
30	34.20	20	241.31	34.17	83	208.19	34.84	333	112.73
50	34.67	34	238.7	34.61	138	201.70	35.45	553	117.45
Ours Net	33.97	82	47.70	34.99	109	35.77	35.02	178	26.31
AG [30]	33.77	487	18.93	34.40	529	16.51	35.15	636	13.08

Table 4. **Ablation of GEM.** Even with 10 principle components a high PSNR of 31.81dB is achieved, while taking only 7MB of memory. In contrast to fixed-sized neural networks, the GEM can be adjusted on the fly depending on the hardware. Moreover, since evaluation requires a single dot product for forward pass the rendering speed is around four times higher than our network. The speed evaluation was done using a single Nvidia A100 GPU.

Figure 9 demonstrates that our method is able to handle different topologies (subject wearing glasses), despite utilizing a fixed UV space.

## 5. Discussion

We design a universal method capable of distilling 3DGS-based avatar solutions into a lightweight representation, GEM, provided that normalized input across training frames is available. The only requirement to successfully distill GEM is to have a dataset with Gaussian-image pairs across the training sequences. Our results show that a compact representation of the linear basis produces state-of-the-art results in terms of quality and speed. Note that

to achieve wrinkle-level details, the generator itself has to produce high-quality outputs. Our distillation technique can be applied to existing methods like [67], making them lightweight and compact. GEM is well-suited for commodity devices, generating Gaussian primitives by a simple linear combination of the basis vectors. This potential has promising implications for tasks like holoportation, audio-driven avatars, and virtual reality.

**Limitations:** The PCA-based GEM models have a global extent which is useful for some applications, but it also means that we cannot control local changes and produce more combinations of local features. Thus, further work could include incorporating a localized PCA basis [37] for better avatar control, which could potentially enable a wider range of expressions outside the training set. Other limitations are; side-view generalization which results in unstable expressions and personalization. For new subjects a new representation has to be learned from multi-view data. An interesting future avenue is to create a statistical model across subjects.

## 6. Conclusion

We have proposed Gaussian Eigen Models for Human Heads, a linear appearance model that represents photo-realistic head avatars. The simplicity of this appearance model results in massively reduced compute requirements in comparison to CNN-based avatar methods. Although the idea is simple, it offers many interesting downstream applications. The lightweight representation could improve the management, sharing, and applicability of avatars. Moreover, GEM simplifies the process of online avatar animation from RGB images and increases flexibility by balancing memory and quality trade-offs through additional control over the number of eigenbases. Our distillation approach can be applied to existing methods, making them available for compression. We demonstrate how GEMs can be used in scenarios like self-reenactment and cross-person animation, even in real-time.

**Acknowledgements** The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting WZ. JT was supported by the ERC Starting Grant LeMo (101162081). All the data were processed outside Google.

## References

- [1] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. PanoHead: Geometry-aware 3D full-head synthesis in 360°. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20950–20959, 2023. 2
- [2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999. 2, 4, 5
- [3] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabriel Schwartz, Michael Zollhoefer, Shunsuke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shou-I Yu, Yaser Sheikh, and Jason M. Saragih. Authentic volumetric avatars from a phone scan. *Transactions on Graphics (TOG)*, 41:1 – 19, 2022. 3
- [4] Eric Chan, Connor Z. Lin, Matthew Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, S. Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16102–16112, 2021. 2
- [5] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20279–20290, 2022. 4, 5, 7
- [6] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhler, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3D morphable face models—past, present, and future. *Transactions on Graphics (TOG)*, 39(5), 2020. 2
- [7] Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, De-jia Xu, and Zhangyang Wang. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps, 2023. 3
- [8] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *Transactions on Graphics (TOG)*, 40:1 – 13, 2020. 2, 5, 6
- [9] Yutao Feng, Xiang Feng, Yintong Shang, Ying Jiang, Chang Yu, Zeshun Zong, Tianjia Shao, Hongzhi Wu, Kun Zhou, Chenfanfu Jiang, and Yin Yang. Gaussian splashing: Dynamic fluid synthesis with gaussian splatting, 2024. 3
- [10] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8645–8654, 2020. 2, 3
- [11] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. *Transactions on Graphics (TOG)*, 41:1 – 12, 2022. 2
- [12] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Monophm: Dynamic head reconstruction from monocular videos. 2024. 5
- [13] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Npga: Neural parametric gaussian avatars, 2024. 3
- [14] Sharath Girish, Kamal Gupta, and Abhinav Shrivastava. Eagles: Efficient accelerated 3d gaussians with lightweight encodings, 2024. 3
- [15] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular RGB videos. pages 18632–18643, 2022. 2, 3, 5
- [16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. 7
- [17] Ying Jiang, Chang Yu, Tianyi Xie, Xuan Li, Yutao Feng, Huamin Wang, Minchen Li, Henry Lau, Feng Gao, Yin Yang, and Chenfanfu Jiang. VR-GS: A physical dynamics-aware interactive gaussian splatting system in virtual reality, 2024. 3
- [18] Ian T. Jolliffe. Principal component analysis and factor analysis. 1986. 4
- [19] Berna Kabadayi, Wojciech Zielonka, Bharat Lal Bhatnagar, Gerard Pons-Moll, and Justus Thies. GAN-Avatar: Controllable personalized gan-based human head avatar. pages 882–892, 2024. 2
- [20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2019. 2
- [21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *Transactions on Graphics (TOG)*, 42:1 – 14, 2023. 2, 3
- [22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 2
- [23] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. NeRSemble: Multi-view radiance field reconstruction of human heads. *Transactions on Graphics (TOG)*, 42:1 – 14, 2023. 5
- [24] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads, 2024. 3
- [25] Christoph Lassner and Michael Zollhofer. Pulsar: Efficient sphere-based neural rendering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1440–1449, 2021. 3
- [26] Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3d gaussian splatting for static and dynamic radiance fields. *arXiv preprint arXiv:2408.03822*, 2024. 3
- [27] Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3d gaussian representation for radiance field. pages 21719–21728, 2024. 3
- [28] Junxuan Li, Shunsuke Saito, Tomas Simon, Stephen Lombardi, Hongdong Li, and Jason M. Saragih. MEGANE: Morphable eyeglass and avatar network. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12769–12779, 2023. 3



- [29] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2, 3, 4
- [30] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4, 5, 6, 7, 8
- [31] Stephen Lombardi, Tomas Simon, Jason M. Saragih, Gabriel Schwartz, Andreas M. Lehrmann, and Yaser Sheikh. Neural volumes. *Transactions on Graphics (TOG)*, 38:1 – 14, 2019. 2, 3
- [32] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason M. Saragih. Mixture of volumetric primitives for efficient neural rendering. *Transactions on Graphics (TOG)*, 40:1 – 13, 2021. 2
- [33] Shugao Ma, Tomas Simon, Jason M. Saragih, Dawei Wang, Yuecheng Li, Fernando De la Torre, and Yaser Sheikh. Pixel codec avatars. pages 64–73, 2021. 4
- [34] Shengjie Ma, Yanlin Weng, Tianjia Shao, and Kun Zhou. 3d gaussian blendshapes for head avatar animation. In *ACM SIGGRAPH Conference Proceedings, Denver, CO, United States, July 28 - August 1, 2024*, 2024. 3
- [35] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, pages 405–421, 2020. 2, 5
- [36] KL Navaneet, Kossar Pourahmadi Meibodi, Soroush Abbasi Koochpayegani, and Hamed Pirsiavash. Compgs: Smaller and faster gaussian splatting with vector quantization. *ECCV*, 2024. 3
- [37] Thomas Neumann, Kiran Varanasi, Stephan Wenger, Markus Wacker, Marcus A. Magnor, and Christian Theobalt. Sparse localized deformation components. *Transactions on Graphics (TOG)*, 32:1 – 10, 2013. 8
- [38] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. ASH: animatable gaussian splats for efficient and photoreal human rendering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1165–1175. IEEE, 2024. 2, 3, 4
- [39] Panagiotis Papantonakis, Georgios Kopanas, Bernhard Kerbl, Alexandre Lanvin, and George Drettakis. Reducing the memory footprint of 3d gaussian splatting. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 7(1):1–17, 2024. 3
- [40] Frederick I. Parke. Computer generated animation of faces. In *Proceedings of the ACM Annual Conference - Volume 1*, page 451–457, New York, NY, USA, 1972. Association for Computing Machinery. 1, 2
- [41] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. *IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, 2009. 3
- [42] Malte Prinzler, Otmar Hilliges, and Justus Thies. DINER: Depth-aware Image-based NEural Radiance fields. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12449–12459, 2022. 2
- [43] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic head avatars with rigged 3D gaussians. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20299–20309, 2024. 2, 3, 5, 6, 7, 8
- [44] Edoardo Remelli, Timur M. Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabián Prada, Jason M. Saragih, and Yaser Sheikh. Drivable volumetric avatars using texel-aligned features. *SIGGRAPH Conference Papers (SA)*, 2022. 3
- [45] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 130–141, 2024. 2, 3, 4
- [46] Robert W. Sumner and Jovan Popović. Deformation transfer for triangle meshes. *SIGGRAPH*, 2004. 4
- [47] Kartik Teotia, R. MallikarjunB., Xingang Pan, Hyeon-Joong Kim, Pablo Garrido, Mohamed A. Elgharib, and Christian Theobalt. HQ3DAvatar: High quality controllable 3D head avatar. *Transactions on Graphics (TOG)*, 43(3):27:1–27:24, 2024. 3
- [48] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B. Goldman, and M. Zollhöfer. State of the art on neural rendering. *EG*, 2020. 2
- [49] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Niessner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhoefer, and Vladislav Golyanik. Advances in neural rendering. *Computer Graphics Forum (CGF)*, pages 703–735, 2022. 2
- [50] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016. 2, 5
- [51] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering. *ACM Transactions on Graphics (TOG)*, 38:1 – 12, 2019. 3
- [52] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 2021. 7
- [53] Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. StyleAvatar: Real-time photo-realistic portrait avatar from a single video. In *SIGGRAPH Conference Papers (SA)*, pages 67:1–67:10, 2023. 2, 3
- [54] Zian Wang, Tianchang Shen, Merlin Nimier-David, Nicholas Sharp, Jun Gao, Alexander Keller, Sanja Fidler, Thomas Müller, and Zan Gojcic. Adaptive shells for efficient neural

- radiance field rendering. *Transactions on Graphics (TOG)*, 42(6), 2023. [3](#)
- [55] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. PhysGaussian: Physics-integrated 3D gaussians for generative dynamics. pages 4389–4398, 2024. [3](#)
- [56] Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Huang Han, Qi Guojun, and Yebin Liu. LatentAvatar: Learning latent expression code for expressive neural head avatar. In *SIGGRAPH Conference Papers (SA)*, pages 86:1–86:10, 2023. [2](#)
- [57] Yuelang Xu, Bengwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian Head Avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1941. IEEE, 2024. [2](#), [3](#), [4](#)
- [58] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. pages 586–595, 2018. [6](#)
- [59] Xiaochen Zhao, Lizhen Wang, Jingxiang Sun, Hongwen Zhang, Jinli Suo, and Yebin Liu. HAvatar: High-fidelity head avatar via facial model conditioned neural radiance field. *Transactions on Graphics (TOG)*, 2023. [2](#)
- [60] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. GPS-gaussian: Generalizable pixel-wise 3D gaussian splatting for real-time human novel view synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19680–19690. IEEE, 2024. [2](#)
- [61] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. GPS-Gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. pages 19680–19690, 2024. [3](#)
- [62] Yufeng Zheng, Victoria Fernández Abrevaya, Xu Chen, Marcel C. Buhler, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13535–13545, 2021. [2](#)
- [63] Yufeng Zheng, Yifan Wang, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. PointAvatar: Deformable point-based head avatars from videos. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21057–21067, 2022.
- [64] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4574–4584, 2022. [2](#), [3](#), [5](#), [6](#), [7](#)
- [65] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)*, 2022. [5](#)
- [66] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3D gaussian avatars. In *International Conference on 3D Vision (3DV)*, 2025. [2](#), [3](#)
- [67] Wojciech Zielonka, Stephan J. Garbin, Alexandros Lattas, George Kopanas, Paulo Gotardo, Thabo Beeler, Justus Thies, and Timo Bolkart. Synthetic prior for few-shot drivable head avatar inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. [3](#), [8](#)
- [68] Michael Zollhöfer, Justus Thies, Darek Bradley, Pablo Garrido, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3D face reconstruction, tracking, and applications. *Computer Graphics Forum (CGF)*, 37(2):523–550, 2018. [2](#)



## A.5 SYNTHETIC PRIOR FOR FEW-SHOT DRIVABLE HEAD AVATAR INVERSION

*Synthetic Prior for Few-Shot Drivable Head Avatar Inversion*

Wojciech Zielonka, Stephan J. Garbin, Alexandros Lattas, George Kopanas, Paulo Gotardo, Thabo Beeler, Justus Thies, Timo Bolkart

Published in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA, 2025.

## Abstract

We present SynShot, a novel method for the few-shot inversion of a drivable head avatar based on a synthetic prior. We tackle three major challenges. First, training a controllable 3D generative network requires a large number of diverse sequences, for which pairs of images and high-quality tracked meshes are not always available. Second, the use of real data is strictly regulated (e.g., under the General Data Protection Regulation, which mandates frequent deletion of models and data to accommodate a situation when a participant’s consent is withdrawn). Synthetic data, free from these constraints, is an appealing alternative. Third, state-of-the-art monocular avatar models struggle to generalize to new views and expressions, lacking a strong prior and often overfitting to a specific viewpoint distribution. Inspired by machine learning models trained solely on synthetic data, we propose a method that learns a prior model from a large dataset of synthetic heads with diverse identities, expressions, and viewpoints. With few input images, SynShot fine-tunes the pretrained synthetic prior to bridge the domain gap, modeling a photorealistic head avatar that generalizes to novel expressions and viewpoints. We model the head avatar using 3D Gaussian splatting and a convolutional encoder-decoder that outputs Gaussian parameters in UV texture space. To account for the different modeling complexities over parts of the head (e.g., skin vs hair), we embed the prior with explicit control for upsampling the number of per-part primitives. Compared to SOTA monocular and GAN-based methods, SynShot significantly improves novel view and expression synthesis.

# Synthetic Prior for Few-Shot Drivable Head Avatar Inversion

Wojciech Zielonka<sup>1, 2, 3\*</sup> Stephan J. Garbin<sup>3</sup> Alexandros Lattas<sup>3</sup>  
George Kopanas<sup>3</sup> Paulo Gotardo<sup>3</sup> Thabo Beeler<sup>3</sup> Justus Thies<sup>1, 2</sup> Timo Bolkart<sup>3</sup>

<sup>1</sup>Max Planck Institute for Intelligent Systems, Tübingen, Germany

<sup>2</sup>Technical University of Darmstadt <sup>3</sup>Google

<https://zielon.github.io/synshot/>

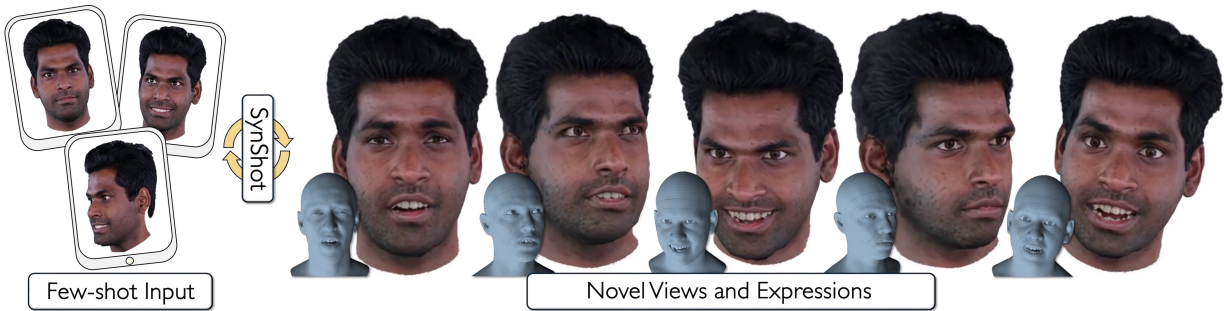


Figure 1. Given a few input images (left), **SynShot** generates a personalized 3D Gaussian avatar that renders from new viewpoints and unseen expressions (right). To compensate for the missing information in the input images, we leverage a generative Gaussian head avatar trained on a diverse synthetic head dataset as a 3D prior.

## Abstract

We present **SynShot**, a novel method for the few-shot inversion of a drivable head avatar based on a synthetic prior. We tackle three major challenges. First, training a controllable 3D generative network requires a large number of diverse sequences, for which pairs of images and high-quality tracked meshes are not always available. Second, the use of real data is strictly regulated (e.g., under the **General Data Protection Regulation**, which mandates frequent deletion of models and data to accommodate a situation when participant’s consent is withdrawn). Synthetic data, free from these constraints, is an appealing alternative. Third, state-of-the-art monocular avatar models struggle to generalize to new views and expressions, lacking a strong prior and often overfitting to a specific viewpoint distribution. Inspired by machine learning models trained solely on synthetic data, we propose a method that learns a prior model from a large dataset of synthetic heads with diverse identities, expressions, and viewpoints. With few input images, **SynShot** fine-tunes the pretrained synthetic prior to bridge the domain gap, modeling a photorealistic head

avatar that generalizes to novel expressions and viewpoints. We model the head avatar using 3D Gaussian splatting and a convolutional encoder-decoder that outputs Gaussian parameters in UV texture space. To account for the different modeling complexities over parts of the head (e.g., skin vs hair), we embed the prior with explicit control for upsampling the number of per-part primitives. Compared to SOTA monocular and GAN-based methods, **SynShot** significantly improves novel view and expression synthesis.

## 1. Introduction

The ability to build high-fidelity drivable digital avatars is a key enabler for virtual reality (VR) and mixed reality (MR) applications. However, creating photorealistic human head models [1, 53] using traditional rendering assets requires sophisticated data capture and significant manual cleanup, which is time-consuming and expensive.

The recent advancements in learning-based methods and radiance fields [29, 43] have simplified the avatar creation process, leading to impressive progress in quality and democratization of neural head avatars [19, 42, 62]. Such progress is particularly noticeable in enhancing control

\*Work done while WZ was interning at Google in Zurich, Switzerland

through lightweight animation [48, 65, 82], and reducing training time to a few minutes [79]. These methods are trained on multi-view [42, 48, 62] or single-view videos [9, 19, 65, 79], typically requiring hundreds to thousands of video frames. Processing such datasets is complex and error-prone as most methods require tracking a coarse head mesh across all frames, which is typically done by fitting a 3D morphable model [38, 45] to the image. A further limitation of existing personalized head avatars is their poor generalization to facial expressions and camera viewpoints not captured in the set of input images.

Another recent body of work addresses the problem of building 3D head avatars from one or few input images, [8, 11, 71]. However, their rendering quality and fidelity are typically lower than those of methods trained on large datasets (e.g., [48, 82]). To improve quality, some methods [69, 70, 75] first learn a multi-identity head model that is used as prior when optimizing for the personalized avatar. Training these head priors requires a large-scale multi-view image dataset that is expensive and time-consuming to capture. Moreover, managing real data under protection laws like **GDPR** is cumbersome for experimentation and maintenance, as users must periodically (e.g., every 30 days) delete all dataset derivatives and trained models, allowing dataset participants to be removed from both if needed. Alternatively, the FFHQ dataset [28] may be employed, with 4D GAN-based methods [13, 56, 74] constructing an inversion prior from it. However, these approaches tend to exhibit artifacts during novel view synthesis and struggle with preserving identity. In summary, the expressive power of this prior is strongly influenced by: the training data diversity (e.g., ethnicity, age, facial features, expressions), the multi-view capture hardware setup (*i.e.*, lighting, view-density, calibration quality, frame-rate), and the quality of the data pre-processing (e.g., mesh tracking, background masking).

In contrast to the previous work that focuses on expensive and cumbersome real data, we overcome these limitations and propose *SynShot*, a new method that builds a prior solely on synthetic data and adapts to a real test subject requiring only a few input images. Building on the success of ML models trained on synthetic data for tasks like 3D face regression [52], 2D landmark prediction [64], rigid face alignment [3], and few-shot head reconstruction [6, 63, 72], *SynShot* is trained solely on a large synthetic dataset generated from 3DMM samples and diverse assets. Synthetic data offers complete control over dataset creation to meet size and diversity needs for training an expressive head prior, eliminating the need for costly capture hardware and addressing privacy concerns with real subjects. The benefits brought by synthetic data come at the cost of having to handle the domain gap between the trained head prior and real images captured “in the wild”. To effectively bridge this gap, we first fit the synthetic prior to real images

and then fine-tune the prior weights to the real data using the pivotal tuning strategy proposed in [49]. With as few as three input images, *SynShot* reconstructs a photorealistic head avatar that generalizes to novel expressions and camera viewpoints (Fig. 1). The results show that our method outperforms state-of-the-art personalized monocular methods [54, 65, 79] trained on thousands of images each. Our method represents head avatars using 3D Gaussian primitives [29], where Gaussian parameters are generated by a convolutional architecture in UV space [39, 50, 82].

In summary, our key contributions are:

1. A generative method based on a convolutional encoder-decoder architecture that is trained on extensive synthetic data only to produce controllable 3D head avatars.
2. A reconstruction scheme that adapts and fine-tunes a pre-trained generative model on a few real images to create a personalized, photorealistic 3D head avatar.

## 2. Related Work

**Few-shot Head Avatars.** 3D Morphable Models (3DMM) [4, 15, 38, 45] have long been used for creating facial avatars. When paired with generative models for textures [22, 35, 36, 41], 3DMMs can be optimized from in-the-wild images. Techniques such as inverse rendering [14], diffusion-based inpainting [44], and pivotal-tuning [37, 49] are used to disentangle appearance from identity. Neural radiance fields (NeRF) [43] and 3D Gaussian representations (3DGS) [29] have also been widely used for avatar reconstruction. EG3D [7] employs features on tri-planes, enabling consistent 3D face generation and inversion from in-the-wild images. PanoHead [2] extends EG3D through tri-grids to achieve a full 360-degree generation of static human heads. Gaussian3Diff [34] further improves quality by replacing neural features with 3D Gaussians. Rodin [63] and RodinHD [72] leverage an extensive dataset of synthetic humans to train a triplane-based avatar generator used to invert in-the-wild images; however, the results remain confined to the synthetic domain and avatars are not drivable.

Diner [46] incorporates depth information, while Preface [5] trains a volumetric prior on synthetic human data and fits it to a few input images to match a subject’s likeness. Cafca [6] extends Preface to better generalize to static but arbitrary facial expressions. In contrast, our method not only bridges the domain gap from synthetic to real but also produces animatable avatars. MofaNeRF [78] and NeRFace [19] condition NeRFs on expression (and shape) codes, while HeadNeRF [25] similarly embeds NeRFs into parametric models. Portrait4D [13] introduces one-shot 4D head synthesis using a transformer-based animatable triplane reconstructor built on the EG3D [7]. Next3D [56] employs GAN-based neural textures embedded on a parametric mesh; however, it suffers from inversion problems. InvertAvatar [74] tackles

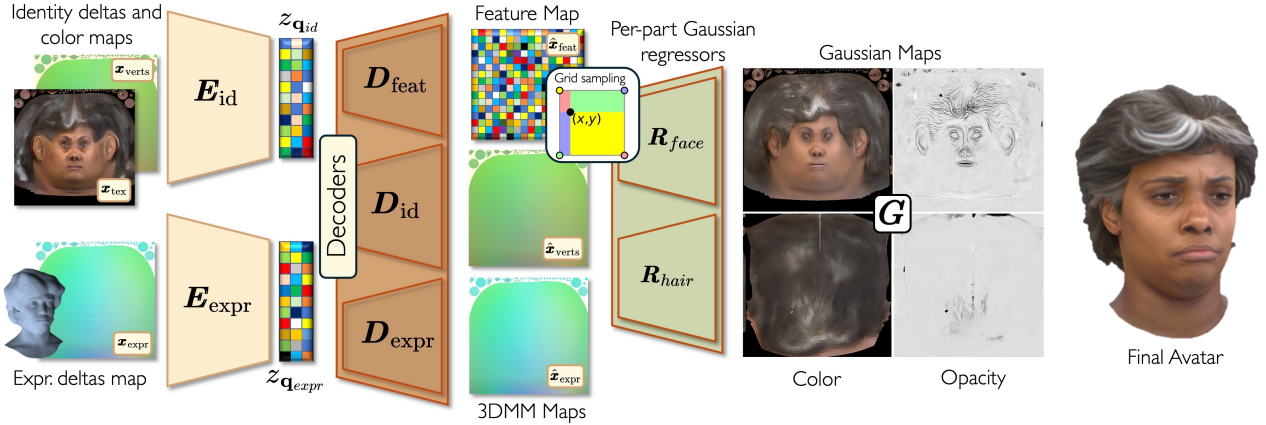


Figure 2. **Pipeline overview.** Given an extracted texture  $\mathbf{x}_{\text{tex}}$ , rasterized position map  $\mathbf{x}_{\text{verts}}$ , and PCA expression deltas  $\mathbf{x}_{\text{exp}}$  our network utilized VQ-VAE to jointly optimize for two latent space  $\mathbf{z}_{\text{expr}}$  and  $\mathbf{z}_{\text{id}}$ . The VQ-VAE decoders predict feature map  $D_{\text{feat}}(\mathbf{q}(\mathbf{z}_{\text{id}}), \mathbf{q}(\mathbf{z}_{\text{expr}})) \rightarrow \hat{\mathbf{x}}_{\text{feat}}$ , identity and color maps  $D_{\text{id}}(\mathbf{q}(\mathbf{z}_{\text{id}})) \rightarrow \{\hat{\mathbf{x}}_{\text{tex}}, \hat{\mathbf{x}}_{\text{verts}}\}$ , and expression deltas  $D_{\text{expr}}(\mathbf{q}(\mathbf{z}_{\text{expr}})) \rightarrow \hat{\mathbf{x}}_{\text{expr}}$ . Finally, bilinearly sampled maps are passed to per-part regressors  $R_{\text{color}}$  and  $R_{\text{gauss}}$  to obtain primitives to rasterize.

the shortcomings of Next3D and further refines avatar inversion using few-shot images. Despite relatively good frontal performance, these GAN-based methods often exhibit artifacts, such as identity changes, in novel view synthesis.

Recent methods by Xu *et al.* [70] are conceptually similar to HeadNeRF and MofaNeRF; however, instead of embedding NeRF [43] on a mesh, they employ 3DGS [29]. GPHM [70] uses a series of MLPs to generate Gaussian primitives attached to a parametric model, enabling expression control and inversion, though it conditions only the avatar’s shape. GPHMv2 [69] extends GPHM with a dynamic module for improved reenactment control and a larger dataset, further enhancing quality. HeadGAP [75] also models avatars using MLPs, utilizing part-based features and additional color conditioning to improve quality. While these methods embed primitives directly on the mesh surface, our approach explicitly learns the primitive parameters by modeling their distribution via a VQ-VAE [61], eliminating the need for a guiding mesh during the test time as the shape is captured within our latent space.

**Multi-view Personalized Avatars.** Volumetric primitives, combined with multi-view training, are highly effective for modeling human heads [23, 27, 31, 50, 57, 58, 68, 82] as they capture intricate details like hair and subsurface scattering [51]. VolTeMorph [21] embeds a NeRF within tetrahedral cages that guide volumetric deformation. Qian *et al.* [48] attach Gaussian primitives to 3DMM triangles, whose local rotations and stretch deform the Gaussians without requiring neural networks. Xu *et al.* [68] and Giebenhain *et al.* [23] extend that work to further predict corrective fields over the Gaussians; rather than colors, they splat features that are translated into color by an image-space CNN [60]. Lombardi *et al.* [40] position 3D voxels with RGB

and opacity values at the vertices of a head mesh, using ray tracing for volumetric integration. Saito *et al.* [50] improve quality by replacing voxel primitives with 3D Gaussians and applying rasterization. Our VQ-GAN training aligns with these principles for *few-shot capture*, as we supervise the process using a hybrid mesh-primitive approach to model the generative distribution.

**Monocular Personalized Avatars.** Monocular methods often rely on a strong 3DMM prior, as recovering a 3D shape from a 2D image is an inherently under-constrained problem. Face2Face [59] was a seminal work that enabled real-time reconstruction and animation of a parametric model. However, it lacks detailed hair representation and relies on low-frequency PCA texture models, which significantly affects quality. This limitation has led to the rise of neural avatars based on NeRF [10, 18–20, 62, 66, 67, 71, 76, 76, 77] and later on 3D Gaussian primitives [9, 32, 54, 65]. INSTA [79] applies triangle deformation gradients [55] to each NeRF sample based on proximity to the nearest triangle, enabling avatar animation. This approach has been adapted to 3D Gaussian Splatting (3DGS) by methods like Flash Avatar [65] or Splatting Avatar [54]. Unlike *SynShot*, these monocular methods do not generalize well to novel views and expressions. Moreover, they require three orders of magnitude more real data to create a avatar. *SynShot* overcomes this by leveraging a synthetic prior during few shot avatar inversion, achieving high-quality results.

### 3. Method

This section describes *SynShot*, how we train the synthetic prior to generate drivable 3D Gaussian head avatars, and how we use it for few-shot head avatar reconstruction.



### 3.1. Preliminaries

We represent a base 3D mesh as  $\mathbf{S} = \bar{\mathbf{S}} + \delta\mathbf{B}_{id} + \gamma\mathbf{B}_{expr}$ , where  $\bar{\mathbf{S}}$  is the average shape,  $\mathbf{B}_{(id,expr)}$  are the bases for identity and expression of a 3DMM, and  $\delta, \gamma$  denote the corresponding coefficients. Additionally, we use linear blend skinning (LBS) for head rotation around the neck with pose corrective offsets, and to rotate the eyeballs.

The head avatar is rendered via 3D Gaussian splatting, using the CUDA implementation of 3DGS [29]. The rasterizer is defined as  $\mathcal{R}(\mathbf{G}, \mathbf{K}) \rightarrow \tilde{\mathcal{I}}$ , for a camera  $\mathbf{K}$  and a set of  $n$  3D Gaussians  $\mathbf{G} \in \mathbb{R}^{n \times (11+16 \times 3)} := \{\phi, \theta, \sigma, \alpha, \mathbf{h}\}$ , with position  $\phi \in \mathbb{R}^{n \times 3}$ , rotation  $\theta \in \mathbb{R}^{n \times 3 \times 3}$ , scale  $\sigma \in \mathbb{R}^{n \times 3}$ , opacity  $\alpha \in \mathbb{R}^n$ , and the (third-degree) spherical harmonics parameters  $\mathbf{h} \in \mathbb{R}^{n \times 16 \times 3}$ , where  $n$  is the number of Gaussians. See Kerbl *et al.* [29] for more details.

### 3.2. Gaussian Prior Model

Our prior is modeled as a generative convolutional network with additional lightweight regressors that output Gaussian 2D maps, i.e. multichannel parameter textures. To sample a flexible number of Gaussian primitives, UV positions and features are bilinearly interpolated from intermediary feature maps, before decoding the standard Gaussian attributes that are rendered using  $\mathcal{R}(\cdot)$ . The architecture of the prior learned by *SynShot* is illustrated in Fig. 2.

**Drivable VQ-VAE.** Our network has an encoder-decoder architecture based on the VQ-VAE [61]. We follow the approach of Esser *et al.* [16], and use a transformer operating in a quantized latent codebook space to better model long-range dependencies between encoded patches in images. The input to the encoder consists of an RGB texture map  $\mathbf{x}_{tex} \in \mathbb{R}^{H \times W \times 3}$ , an XYZ vertex position map  $\mathbf{x}_{verts} = \mathcal{R}_{uv}(\delta\mathbf{B}_{id}) \in \mathbb{R}^{H \times W \times 3}$  representing the rasterized positions of the neutral mesh, and an expression map  $\mathbf{x}_{exp} = \mathcal{R}_{uv}(\gamma\mathbf{B}_{expr}) \in \mathbb{R}^{H \times W \times 3}$  denoting rasterized expression offsets from the neutral mesh, where  $\mathcal{R}_{uv}(\cdot)$  denotes UV space rasterization. The encoder network consists of two parallel branches, one for identity and one for expression. This way we explicitly disentangle static components, such as face shape and appearance, from dynamic ones, such as wrinkles, and self-shadowing using two separate latent spaces. We denote them as  $\mathbf{E}_{id}(\mathbf{x}_{tex}, \mathbf{x}_{verts}) \rightarrow \mathbf{z}_{id}$ , where  $\mathbf{z}_{id} \in \mathbb{R}^{h \times w \times n_{id}}$  is the identity code and  $\mathbf{E}_{expr}(\mathbf{x}_{exp}) \rightarrow \mathbf{z}_{expr}$  with  $\mathbf{z}_{expr} \in \mathbb{R}^{h \times w \times n_{expr}}$  representing the expression code. The identity and expression latents undergo element-wise quantization  $\mathbf{q}(\cdot)$ . For simplicity, we omit the subscript and let  $\mathbf{z} \in \{\mathbf{z}_{id}, \mathbf{z}_{expr}\}$  denote identity and expression latent codes, with spatial codes  $z_{ij} \in \mathbb{R}^n$ , which we quantize by:

$$\mathbf{q}(\mathbf{z}) := \left( \arg \min_{z_k \in \mathcal{Z}} \|\mathbf{z}_{ij} - z_k\| \right), \quad (1)$$

with a learned discrete codebook  $\mathcal{Z} = \{z_k\}_{k=1}^K$ , with

$z_k \in \mathbb{R}^n$ . The quantized latent codes are fed into the decoder, which is implemented as three output branches: a feature map decoder,  $\mathbf{D}_{feat}(\mathbf{q}(\mathbf{z}_{id}), \mathbf{q}(\mathbf{z}_{expr})) \rightarrow \hat{\mathbf{x}}_{feat} \in \mathbb{R}^{H \times W \times F}$  with  $F$ -dimensional feature vectors per texel; an identity map decoder,  $\mathbf{D}_{id}(\mathbf{q}(\mathbf{z}_{id})) \rightarrow \{\hat{\mathbf{x}}_{tex}, \hat{\mathbf{x}}_{verts}\}$ ; and an expression decoder,  $\mathbf{D}_{expr}(\mathbf{q}(\mathbf{z}_{expr})) \rightarrow \hat{\mathbf{x}}_{expr}$ . Given the output vertex position and expression maps,  $\hat{\mathbf{x}}_{verts}$  and  $\hat{\mathbf{x}}_{expr}$ , the positions of the Gaussian primitives are then computed as  $\phi = \hat{\mathbf{x}}_{verts} + \hat{\mathbf{x}}_{expr}$ .

**Gaussian Primitives Regression.** A common limitation of using CNNs to regress Gaussian maps is the fixed output resolution, which ties the number of primitives to the output dimensions. This restriction can significantly limit the quality of the reconstructed avatar (see Table 1). To address this issue, we use a part-based densification mechanism. Similar to Kirschstein *et al.* [32], we use bilinear sampling,  $\mathcal{B}(\cdot, u, v)$  to sample the output of the decoders at UV-positions  $(u, v)$ . As different head regions  $r \in \{face, hair\}$  have varying requirements for the density of Gaussian primitives, we bilinearly sample separate parameter maps for the face and scalp region, rather than a single joint map. Thus, per-part map sampling acts as adaptive primitive densification for the individual regions to improve visual quality (Table 1).

We define the primitive positions in the 3DMM space using only shape and expression. Global rotation, translation, and linear blend skinning (LBS) are factored out and applied to the primitives just before splatting to place them in the correct world space. We compute initial *per-part Gaussian parameters* for our primitives. Note that we do not use a fixed canonical space [23, 32, 82], as our initialization is derived from predicted position maps. We first obtain positions by sampling  $\phi_r = \mathcal{B}(\phi, u_r, v_r)$ , for  $r \in \{face, hair\}$ . Next, for each  $\phi_r$ , we compute nearest neighbor distance and initialize scale as  $\sigma_r = \min_{j \neq i} \|\phi_{r_i} - \phi_{r_j}\|_2$ . Initial opacity is set to  $\alpha = 0.7$ . Finally, the per-part rotations are computed as  $\theta_r = \begin{bmatrix} \frac{\mathbf{T}}{\|\mathbf{T}\|} & \frac{\mathbf{B}}{\|\mathbf{B}\|} & \frac{\mathbf{N}}{\|\mathbf{N}\|} \end{bmatrix} \in \mathbb{R}^{h \times w \times 3 \times 3}$ , based on the image space gradient:

$$\mathbf{T} = \frac{\partial \phi_r}{\partial u}, \quad \mathbf{B} = \frac{\partial \phi_r}{\partial v}, \quad \mathbf{N} = \mathbf{T} \times \mathbf{B}. \quad (2)$$

Following common practice [23, 32, 50, 68, 75, 81, 82], we predict a neural corrective field for all Gaussian parameters. For this, we use the regressed feature map  $\hat{\mathbf{x}}_{feat}$ , sampling  $\mathbf{s}_r = \mathcal{B}(\hat{\mathbf{x}}_{feat}, u_r, v_r)$ , and lightweight regressors composed of four stacked convolutional blocks with skip connections. Per region, we define two regressors:

$$\mathbf{R}_{color}(\mathbf{s}_r) \rightarrow \mathbf{h}_r \in \mathbb{R}^{h \times w \times 16 \times 3}, \quad (3)$$

$$\mathbf{R}_{gauss}(\mathbf{s}_r) \rightarrow \{\delta\phi_r, \delta\theta_r, \delta\sigma_r, \delta\alpha_r\}, \quad (4)$$

where  $\mathbf{R}_{color}$  regresses the spherical harmonics coefficients  $\mathbf{h}_r$ , and  $\mathbf{R}_{gauss}$  regresses additive parameter offsets

$\Delta := \{\delta\phi_r, \delta\theta_r, \delta\sigma_r, \delta\alpha_r\}$  from the per-part Gaussian parameters. Finally, we apply  $\Delta$  to the primitives of the individual parts, concatenate them, and splat as  $\mathcal{R}(\mathbf{G}, \mathbf{K}) \rightarrow \bar{\mathcal{I}}$ , where  $\bar{\mathcal{I}}$  is the final rendered image and  $\mathbf{G}$  represents the combined Gaussian primitives.

**Training Objectives.** We supervise the training of our model by minimizing the photometric loss:

$$\mathcal{L}_{\text{color}} = \alpha\mathcal{L}_{L1} + \beta\mathcal{L}_{\text{SSIM}} + \gamma\mathcal{L}_{\text{LPIPS}} \quad (5)$$

between the pairs of input and output maps  $\{\mathbf{x}_{\text{tex}}, \hat{\mathbf{x}}_{\text{tex}}\}$ ,  $\{\mathbf{x}_{\text{verts}}, \hat{\mathbf{x}}_{\text{verts}}\}$ , and  $\{\mathbf{x}_{\text{exp}}, \hat{\mathbf{x}}_{\text{exp}}\}$ , and between the pairs of target images and the final splatted images  $\{\mathcal{I}, \bar{\mathcal{I}}\}$ .

Additionally, the position maps  $\hat{\mathbf{x}}_{\text{verts}}$  and expression maps  $\hat{\mathbf{x}}_{\text{exp}}$  are supervised by  $\mathcal{L}_{\text{geom}} = \delta\mathcal{L}_{L1}$ . The final loss is defined as  $\mathcal{L} = \mathcal{L}_{\text{color}} + \mathcal{L}_{\text{geom}}$ . Moreover, we apply  $L_2$  regularization on position, scale, opacity, and the FC ( $l \geq 1$ ) part of the spherical harmonics coefficients: The final loss is defined as  $\mathcal{L} = \mathcal{L}_{\text{photo}} + \mathcal{L}_{\text{reg}}$ . We train our network end-to-end using 8 GPUs Nvidia H100 with batch size 16 (2 per GPU). We optimize the network for 500K iterations with the Adam optimizer [30] with  $\text{lr}=1.3e^{-5}$  and multi-step scheduler which decays the learning rate every milestone by  $\text{gamma}=0.66$ .

### 3.3. Few-shot Avatar Reconstruction

To bridge the gap between in-the-wild and synthetic avatars, we carefully designed a two-stage inversion process based on pivotal fine-tuning [49]. First, we optimize the encoder  $\mathbf{E}_{\text{id}}$  while keeping the rest of the network fixed such that we recover  $\mathbf{z}_{\text{id}}$ . Note that  $\mathbf{E}_{\text{expr}}$  remains unchanged as it should model independent expressions. Once  $\mathbf{E}_{\text{id}}$  is fine-tuned, we fix its predicted identity latent code  $\mathbf{z}_{\text{id}}$ , we fine-tune the decoders  $\{\mathbf{D}_{\text{feat}}, \mathbf{D}_{\text{id}}, \mathbf{D}_{\text{expr}}\}$  and the regressors  $\{\mathbf{R}_{\text{color}}, \mathbf{R}_{\text{gauss}}\}$  for the hair and face regions (Fig. 3).

To make the problem tractable, we employ a few heuristics to aid the optimization. These include early stopping with a warmup phase and an exponential moving average on the loss to determine the stopping criteria. Additionally, we scale the number of optimization steps based on the number of training frames, using a constant factor of 10 to increase the likelihood that each sample is seen at least once. As a training objective, in addition to our photometric term  $\mathcal{L}_{\text{color}}$  (Eq. 5), we follow Lattas *et al.* [37] and, based on ArcFace [12], define two additional objectives:  $\mathcal{L}_{\text{id}}$  and  $\mathcal{L}_{\text{arc}}$ . The final inversion loss is equal to  $\mathcal{L} = \mathcal{L}_{\text{color}} + \mathcal{L}_{\text{arc}} + \mathcal{L}_{\text{id}}$ . For a number of views, up to 20, the optimization takes less than 10 minutes on a single Nvidia H100 which is comparable to INSTA [79]. The training time increases with the number of frames as we scale the iterations accordingly.

### 3.4. Synthetic Dataset

Our dataset consists of approximately 2,000 unique identities, which we render with resolution  $768 \times 512$  using

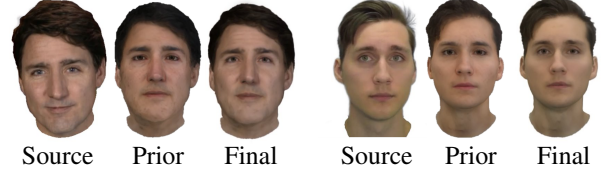


Figure 3. Result of the pivotal tuning before (Prior) and after fine-tuning the model decoders and regressors (Final).

Blender (Cycles) following Wood *et al.* [64], see Fig. 4. We positioned fourteen cameras in front of the subject and an additional fourteen cameras sampled from the upper hemisphere, centered on the scene. We randomly assign assets such as hairstyles and beards to these avatars. Additionally, we utilize high-quality face textures which are randomly distributed among the samples. By combining different shapes and appearances, we augment the set of identities, following practices in synthetic data [64] and 3D face reconstruction [14, 37, 44]. To incorporate tracked expressions from multi-view setups, we propagate them to the avatars during sequence rendering. We additionally compute a hair proxy from strands by voxelizing and fitting it to the scalp region; we apply the same approach for beards. Using a neutral mesh and its hair proxy, we backproject the images onto the texture map. During test time, we use a 3DMM regressor and the input images to extract a texture, which is then used as an initialization for our method. In total, our dataset comprises 14 million images.



Figure 4. Random samples of our synthetic dataset show a diverse range of identities, expressions, and hairstyles that would be challenging to capture in an in-house studio with real subjects.

## 4. Results

We compare *SynShot* to two different types of methods, state-of-the-art personalized monocular methods, and inversion-based general methods. The personalized monocular methods are controlled by FLAME [38] meshes and include INSTA [79], Flash Avatar [65], and SplattingAvatar [54]. For monocular methods, we used an ensemble of four datasets [19, 24, 76, 82] processed using the face tracker from Zielonka *et al.* [80]. SplattingAvatar follows the approach of Zheng *et al.* [76] and uses the monocular 3D face regressor DECA [17] for tracking. In our experiments, we adopted a similar approach, employing an in-house regressor, similar to DECA, to estimate 3DMM expression and pose parameters. While these methods produce photorealistic



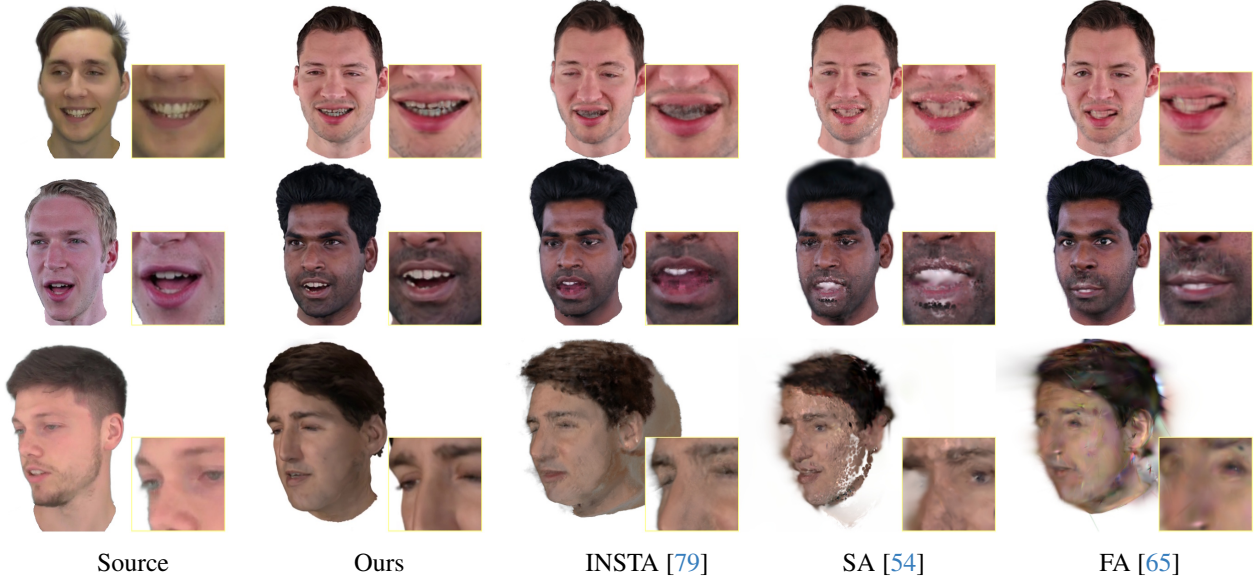


Figure 5. Cross-reenactment comparison of *SynShot* inversion using only 3 views to state-of-the-art (SOTA) methods: **INSTA** [79], Flash Avatar (**FA**) [65], and Splatting Avatar (**SA**) [54], each of which was trained on an average of 3000 frames. It is evident that without a strong prior, these methods fail to generalize to novel expressions and views. Inversion input images are in the supplemental materials.



Figure 6. Novel view evaluation of long hair and beard inversion using only three input images demonstrates the strong generalization capability of *SynShot*.

tic avatars, they struggle with generalization to novel views and poses (see Figure 5). For inversion-based methods, we compare PanoHead [2], HeadNeRF [25], and MofaNeRF [78]. We also compare to concurrent works including Portrait4D [13], Next3D [56], and InvertAvatar [74] (Figures 8 and 9). We use three images for all inversion experiments see supp. material. Figure 6 presents a novel view evaluation of challenging long hair and beard inversion, demonstrating the generalization capabilities of *SynShot*.

**Evaluation.** To measure the performance of *SynShot* without introducing bias, we selected training frames from  $\{F_n\}_{n=1}^{16} = \{1, \dots, 987\}$ , where  $F_n$  denotes the Fibonacci sequence. For all experiments, we use progressive farthest point sampling [47] in the 3DMM expression space to select a specified number of frames from the training set. The self-reenactment sequences were evaluated using LPIPS and SSIM on the last 600 frames from the INSTA dataset [79].

**Monocular Avatar Self-Reenactment.** Our combined

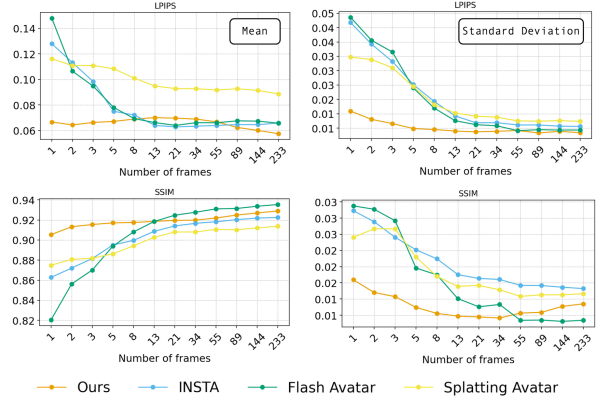


Figure 7. We evaluated the reconstruction error with respect to the number of frames using LPIPS and SSIM metrics. For each frame count, we report the average error (left) and standard deviation (right) over 600 frames across 11 subjects, highlighting the importance of our synthetic prior.

dataset consists of eleven monocular sequences ( $512 \times 512$  resolution), many of which are in-the-wild videos with very limited head motion, resulting in a low error as the test sequences closely resemble the training data, leaving limited room to assess diversity. To address this and accurately measure the effective error, we trained each method on a varying number of frames, corresponding to frames used in our inversion pipeline. The reconstruction error is evaluated on 600 test frames. Figure 7 demonstrates the effective-

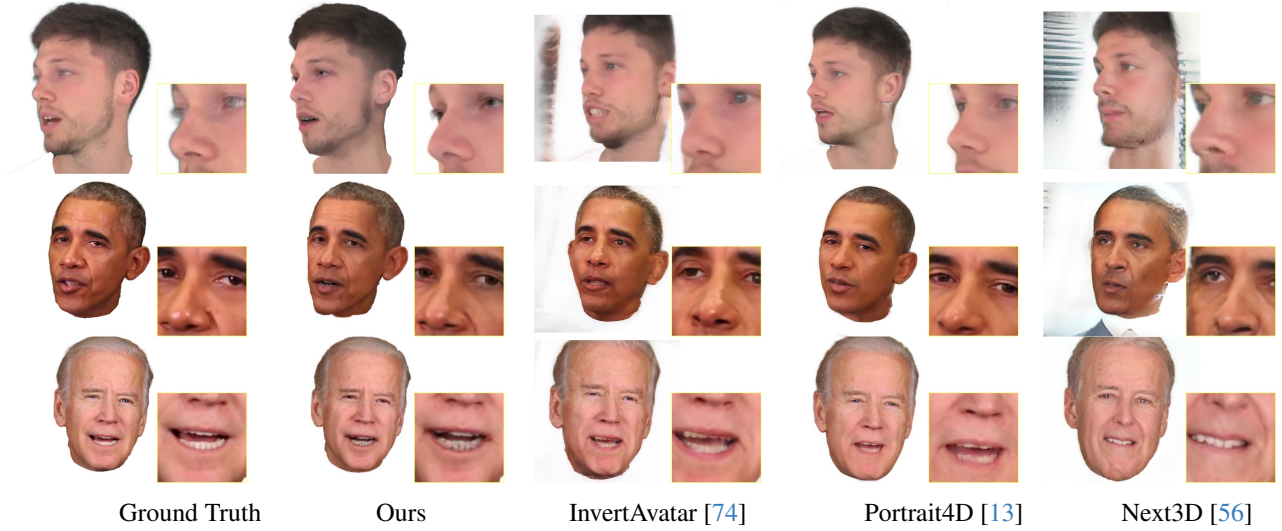


Figure 8. The GAN-based self-reenactment comparison again shows that SynShot better captures identity and synthesizes novel views, proving its usefulness as a synthetic prior and for pivotal fine-tuning in inversion. LPIPS scores: Ours (**0.0236**), InvertAvatar (0.0962), Portrait4D (0.0843), and Next3D (0.2274). Inversion input images can be found in supplemental materials.

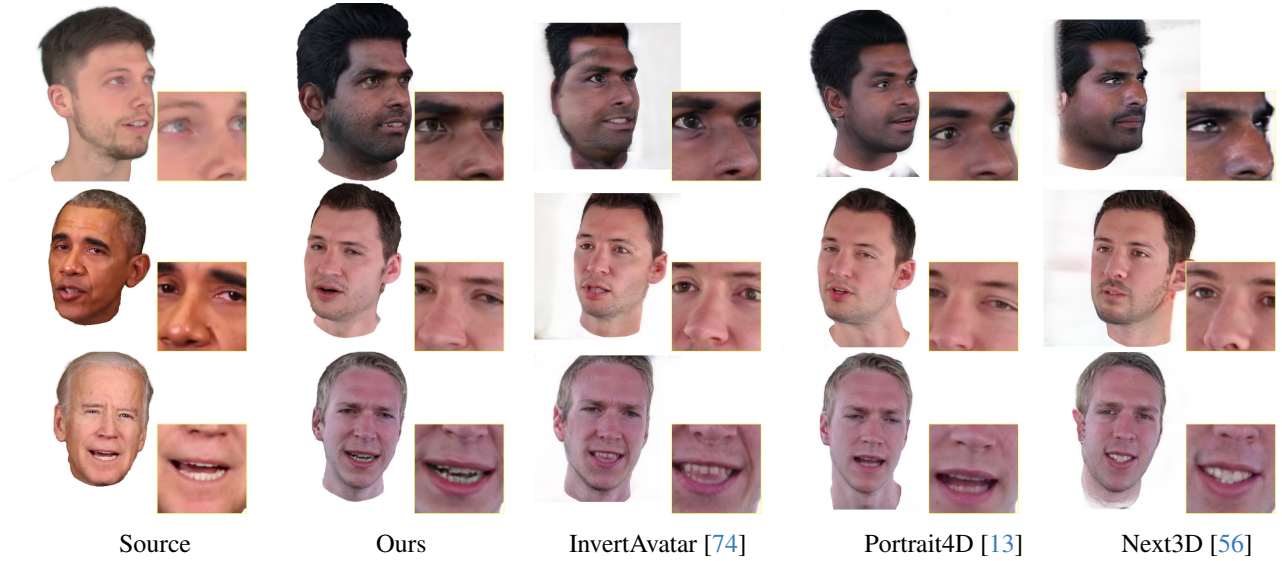


Figure 9. The GAN-based cross-reenactment comparison shows that our method better reconstructs the target subject’s appearance (identity) and remains faithful to the source subject’s head poses and expressions, whereas the other methods suffer from artifacts.

ness of our inversion, particularly with up to 233 training frames. Due to the lack of a strong prior, monocular methods fail in low training frame regimes, and, even with larger training datasets, they do not perform well and produce artifacts. Please note that to benefit from an increased number of input frames, i.e., to ground the avatar reconstruction more on the input than the synthetic prior, it requires an increased number of optimization iterations during pivotal fine-tuning. The number of iteration steps affects the metrics, causing LPIPS to vary non-monotonically.

**Monocular Avatar Cross-Reenactment.** We would like to emphasize the importance of evaluating cross-reenactment, which often reveals issues with generalization and overfitting; however, these aspects are frequently underemphasized, as evaluation sequences are commonly not sufficiently challenging. For instance, Figure 7 indicates that 13 frames may be sufficient for monocular methods to perform well on the test set. Despite achieving high-quality results, most monocular methods [19, 24, 54, 65, 79] struggle with cross-reenactment involving novel expressions and

views. In the supp. mat. we present a full evaluation. Without a strong prior, these methods frequently exhibit artifacts when driven by out-of-distribution sequences. In contrast, our method, leveraging only three images and a synthetic prior with effective shape-expression disentanglement, is able to invert an avatar that significantly outperforms state-of-the-art models trained on thousands of frames. Figure 5 demonstrates cross-reenactment, with the leftmost column serving as the source for expression and view. This shows that incorporating a strong prior enhances the visual quality.

Architecture	L1 ↓	LPIPS ↓	SSIM ↑	PSNR ↑
$F = 128$	0.0356	<b>0.2686</b>	0.8189	20.1536
Tex. up-sampling	<b>0.0352</b>	0.2695	<b>0.8196</b>	<b>20.1909</b>
Single Layer	0.0369	0.2702	0.8177	19.8871
$F = 32$	0.0375	0.2732	0.8146	19.7002
w/o VQ	0.0396	0.2747	0.8122	19.2861
$F = 64$	0.0400	0.2765	0.8104	19.2731
No Sampling	0.0403	0.2853	0.8158	19.9787
$256 \times 256$	0.0365	0.2865	0.8194	20.4010

Table 1. We evaluated various configurations of our VQ-VAE. Each configuration uses the final textures of  $512 \times 512$ , unless stated otherwise. As our final model ( $F = 128$ ) we selected the one which produces sharpest results in terms of LPIPS.

**GAN-based baselines.** We compared SynShot to three animatable GAN-based methods. For our method and InvertAvatar [74], we used three input images, whereas Portrait4D [13] and Next3D [56] are single-shot. Figure 8 presents qualitative self-reenactment results, with additional quantitative LPIPS scores: Ours (**0.0236**), InvertAvatar (0.0962), Portrait4D (0.0843), and Next3D (0.2274). Both results show that SynShot significantly outperforms the baselines. Moreover, Figure 9 presents expression transfer, where our method best captures the subject’s identity and is more stable for novel views and expressions, whereas GAN-based methods tend to introduce artifacts in side views.

**VQ-VAE Architecture Ablation.** Table 1 presents an ablation study of our VQ-VAE architecture. Each model was evaluated on 50 test actors excluded from the training set. Our best model, in terms of sharpness and quality, regresses a feature map  $\hat{\mathbf{x}}_{\text{feat}} \in \mathbb{R}^{H \times W \times F}$ , where  $F = 128$ , at a resolution of  $512 \times 512$ . Regressing Gaussian primitives directly (*No Sampling*) suffers from lack of quality. Using a *Single Layer* instead of two (for hair + face) results in a lower number of Gaussians, which also decreases the final quality. A key feature of our network is densification through texture sampling. In the (*Tex. up-sampling*) experiment, we predict feature maps at  $256 \times 256$  resolution compared to  $512 \times 512$  and apply bilinear sampling to upscale the per-region sampled feature maps to  $512 \times 512$ . This approach achieves results that are almost on par while saving VQ-VAE computation and memory. Finally, using codebook quantization

of latent space improves the final image quality (w/o VQ).

Loss	L1 ↓	LPIPS ↓	SSIM ↑	PSNR ↑
$\mathcal{L}_{\text{photo}} + \mathcal{L}_{\text{VGG}} + \mathcal{L}_{\text{Id}} + \mathcal{L}_{\text{ArcFeat}}$	0.0229	0.0776	0.9073	23.7474
$\mathcal{L}_{\text{photo}} + \mathcal{L}_{\text{VGG}}$	0.0244	0.0839	0.9058	23.1191
$\mathcal{L}_{\text{photo}} + \mathcal{L}_{\text{VGG}} + \mathcal{L}_{\text{Id}}$	0.0246	0.0848	0.9048	23.1949
$\mathcal{L}_{\text{photo}} = \mathcal{L}_{\text{L1}} + \mathcal{L}_{\text{SSIM}}$	0.0217	0.0904	0.9094	23.7331

Table 2. Ablation for our inversion losses.

**Inversion Ablation.** Our inversion pipeline consists of several losses that help bridge the gap between synthetic and real images. This is an important step in our pipeline, as real subjects often have appearance and illumination conditions that differ significantly from our distribution. To address this, we rely on pixel-wise losses and, most importantly, on perceptual losses, which have been shown to aid in effectively matching two distributions [5, 6, 26, 37, 73]. Table 2 shows the inversion reconstruction error using different combinations of losses. As can be seen, using only  $\mathcal{L}_{\text{photo}}$  is insufficient. The combination of  $\mathcal{L}_{\text{VGG}}$ , based on AlexNet [33],  $\mathcal{L}_{\text{ID}}$ , and  $\mathcal{L}_{\text{ArcFeat}}$  provides the best results.

## 5. Discussion

While significantly outperforming monocular methods, *SynShot* has certain limitations that we identify. A key challenge is bridging the domain gap between synthetic and real data. There is considerable room for improvement in the generation of synthetic data. For example, all our synthetic subjects share the same teeth geometry and texture. As a consequence, teeth in our inverted head avatars often closely follow the prior and do not adapt easily. Furthermore, our synthetic data lacks diverse expression-dependent wrinkles, affecting its overall visual quality. Additionally, our dataset was ray-traced with a single environment map, limiting generalization to varied lighting conditions.

## 6. Conclusion

We have proposed *SynShot*, a method for reconstructing a personalized 3D Gaussian head avatar from just a few images. *SynShot* builds a generative head avatar purely from synthetic data and then utilizes this model as a prior in an inversion pipeline. This inversion pipeline follows a pivotal tuning strategy that successfully bridges the domain gap between the prior and the real input images. We demonstrate that our personalized head avatars generalize better to unseen expressions and viewpoints than SOTA head avatars.

**Acknowledgement** The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting WZ. JT was supported by the ERC Starting Grant LeMo (101162081). We also would like to thank Daoye Wang for assisting with synthetic asset generation, Mark Murphy for his help with using Google infrastructure, and Menglei Chai for providing the hair proxy.



## References

- [1] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. The digital emily project: photoreal facial modeling and animation. In *ACM SIGGRAPH 2009 Courses*. Association for Computing Machinery, 2009. [1](#)
- [2] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. PanoHead: Geometry-aware 3D full-head synthesis in 360deg. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20950–20959, 2023. [2](#), [6](#)
- [3] Jan Bednarík, Erroll Wood, Vasileios Choutas, Timo Bolkart, Daoye Wang, Chenglei Wu, and Thabo Beeler. Learning to stabilize faces. *Computer Graphics Forum (CGF)*, 43(2), 2024. [2](#)
- [4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999. [2](#)
- [5] Marcel C. Buehler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, and Abhimitra Meka. Preface: A data-driven volumetric prior for few-shot ultra high-resolution face synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. [2](#), [8](#)
- [6] Marcel C. Buehler, Gengyan Li, Erroll Wood, Leonhard Helminger, Xu Chen, Tanmay Shah, Daoye Wang, Stephan Garbin, Sergio Orts-Escolano, Otmar Hilliges, Dmitry Lagun, Jérémy Riviere, Paulo Gotardo, Thabo Beeler, Abhimitra Meka, and Kripasindhu Sarkar. Cafca: High-quality novel view synthesis of expressive faces from casual few-shot captures. In *SIGGRAPH Asia Conference Papers (SA)*, 2024. [2](#), [8](#)
- [7] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, 2022. [2](#)
- [8] Xiyi Chen, Marko Mihajlovic, Shaofei Wang, Sergey Prokudin, and Siyu Tang. Morphable diffusion: 3D-consistent diffusion for single-image avatar creation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10359–10370. IEEE, 2024. [2](#)
- [9] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Mono-GaussianAvatar: Monocular gaussian point-based head avatar. In *SIGGRAPH Conference Papers (SA)*, page 58. ACM, 2024. [2](#), [3](#)
- [10] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [3](#)
- [11] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar, 2024. [2](#)
- [12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. [5](#)
- [13] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. [2](#), [6](#), [7](#), [8](#)
- [14] Abdallah Dib, Luiz Gustavo Hafemann, Emeline Got, Trevor Anderson, Amin Fadaeinejad, Rafael MO Cruz, and Marc-André Carbonneau. MoSAR: Monocular semi-supervised model for avatar reconstruction using differentiable shading. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1770–1780, 2024. [2](#), [5](#)
- [15] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3D morphable face models—past, present and future. *Transactions on Graphics (TOG)*, 39(5):157:1–157:38, 2020. [2](#)
- [16] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. [4](#)
- [17] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *Transactions on Graphics, (Proc. SIGGRAPH)*, 40(4):88:1–88:13, 2021. [5](#)
- [18] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J. Black. Learning disentangled avatars with hybrid 3D representations, 2023. [3](#)
- [19] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2021. [1](#), [2](#), [5](#), [7](#)
- [20] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. *Transactions on Graphics (TOG)*, 41(6):1–12, 2022. [3](#)
- [21] Stephan Garbin, Marek Kowalski, Virginia Estellers, Stanislaw Szymanowicz, Shideh Rezaeifar, Jingjing Shen, Matthew Johnson, and Julien Valentin. VolTeMorph: Real-time, controllable and generalizable animation of volumetric representations. *Computer Graphics Forum*, 43, 2024. [3](#)
- [22] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1164, 2019. [2](#)
- [23] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. NPGA: Neural parametric gaussian avatars. In *SIGGRAPH Conference Papers (SA)*, 2024. [3](#), [4](#)
- [24] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular RGB videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18653–18664, 2022. [5](#), [7](#)
- [25] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. HeadNeRF: A real-time nerf-based parametric

- head model. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6
- [26] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016. 8
- [27] Berna Kabadayi, Wojciech Zielonka, Bharat Lal Bhatnagar, Gerard Pons-Moll, and Justus Thies. GAN-Avatar: Controllable personalized GAN-based human head avatar. In *International Conference on 3D Vision (3DV)*, 2024. 3
- [28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
- [29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *Transactions on Graphics (TOG)*, 42(4), 2023. 1, 2, 3, 4
- [30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015. 5
- [31] Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. DiffusionAvatars: Deferred diffusion for high-fidelity 3D head avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5492, 2024. 3
- [32] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads. *arXiv preprint arXiv:2406.09377*, 2024. 3, 4
- [33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012. 8
- [34] Yushi Lan, Feitong Tan, Di Qiu, Qiangeng Xu, Kyle Genova, Zeng Huang, Sean Fanello, Rohit Pandey, Thomas Funkhouser, Chen Change Loy, et al. Gaussian3diff: 3d gaussian diffusion for 3d full head synthesis and editing. *arXiv preprint arXiv:2312.03763*, 2023. 2
- [35] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. AvatarMe: Realistically renderable 3D facial reconstruction “in-the-wild”. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 757–766, 2020. 2
- [36] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P Zafeiriou. AvatarMe++: Facial shape and BRDF inference with photorealistic rendering-aware GANs. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 2
- [37] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. FitMe: Deep photorealistic 3D morphable model avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8629–8640, 2023. 2, 5, 8
- [38] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2, 5
- [39] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable Gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [40] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *Transactions on Graphics (TOG)*, 40(4), 2021. 3
- [41] Huiwen Luo, Koki Nagano, Han-Wei Kung, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, and Hao Li. Normalized avatar synthesis using stylegan and perceptual refinement. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11662–11672, 2021. 2
- [42] Shugao Ma, Tomas Simon, Jason M. Saragih, Dawei Wang, Yuecheng Li, Fernando De la Torre, and Yaser Sheikh. Pixel codec avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 64–73. Computer Vision Foundation / IEEE, 2021. 1, 2
- [43] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020. 1, 2, 3
- [44] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, and Stefanos Zafeiriou. Relightify: Relightable 3D faces from a single image via diffusion models. In *International Conference on Computer Vision (ICCV)*, pages 8806–8817, 2023. 2, 5
- [45] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. *IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, 2009. 2
- [46] Malte Prinzler, Otmar Hilliges, and Justus Thies. Diner: Depth-aware image-based neural radiance fields. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [47] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, 2017. 6
- [48] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic head avatars with rigged 3D gaussians. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20299–20309, 2024. 2, 3
- [49] Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *Transactions on Graphics (TOG)*, 42(1), 2022. 2, 5
- [50] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 130–141, 2024. 2, 3, 4
- [51] Kripasindhu Sarkar, Marcel C. Böhler, Gengyan Li, Daoye Wang, Delio Vicini, Jérémy Riviere, Yinda Zhang, Sergio Orts-Escolano, Paulo F. U. Gotardo, Thabo Beeler, and Abhimitra Meka. LitNeRF: Intrinsic radiance decomposition for high-quality view synthesis and relighting of faces.

- In *SIGGRAPH Asia Conference Papers (SA)*, pages 42:1–42:11. ACM, 2023. 3
- [52] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *International Conference on Computer Vision (ICCV)*, pages 1585–1594. IEEE Computer Society, 2017. 2
- [53] Mike Seymour, Chris Evans, and Kim Libreri. Meet mike: epic avatars. In *ACM SIGGRAPH 2017 VR Village*. Association for Computing Machinery, 2017. 1
- [54] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 5, 6, 7
- [55] Robert W. Sumner and Jovan Popović. Deformation transfer for triangle meshes. *ACM SIGGRAPH*, 2004. 3
- [56] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *CVPR*, 2023. 2, 6, 7, 8
- [57] Kartik Teotia, Hyeonwoo Kim, Pablo Garrido, Marc Habermann, Mohamed Elgharib, and Christian Theobalt. GaussianHeads: End-to-end learning of drivable gaussian head avatars from coarse-to-fine representations, 2024. 3
- [58] Kartik Teotia, Mallikarjun B R, Xingang Pan, Hyeonwoo Kim, Pablo Garrido, Mohamed Elgharib, and Christian Theobalt. HQ3DAvatar: High-quality implicit 3D head avatar. *Transactions on Graphics (TOG)*, 43(3):27:1–27:24, 2024. 3
- [59] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3
- [60] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Transactions on Graphics (TOG)*, 38(4):66:1–66:12, 2019. 3
- [61] Aaron Van Den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3, 4
- [62] Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. StyleAvatar: Real-time photo-realistic portrait avatar from a single video. In *SIGGRAPH Conference Papers (SA)*, 2023. 1, 2, 3
- [63] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4563–4573, 2023. 2
- [64] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *International Conference on Computer Vision (ICCV)*, pages 3681–3691, 2021. 2, 5
- [65] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. FlashAvatar: High-fidelity head avatar with efficient gaussian embedding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 5, 6, 7
- [66] Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. AvatarMAV: Fast 3D head avatar reconstruction using motion-aware neural voxels. In *SIGGRAPH Conference Papers (SA)*, pages 1–10, 2023. 3
- [67] Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Huang Han, Qi Guojun, and Yebin Liu. LatentAvatar: Learning latent expression code for expressive neural head avatar. In *SIGGRAPH Conference Papers (SA)*, 2023. 3
- [68] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian Head Avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 4
- [69] Yuelang Xu, Zhaoqi Su, Qingyao Wu, and Yebin Liu. Gphm: Gaussian parametric head model for monocular head avatar reconstruction. *arXiv preprint arXiv:2407.15070*, 2024. 2, 3
- [70] Yuelang Xu, Lizhen Wang, Zerong Zheng, Zhaoqi Su, and Yebin Liu. 3D gaussian parametric head model. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3
- [71] Zhixuan Yu, Ziqian Bai, Abhimitra Meka, Feitong Tan, Qiangeng Xu, Rohit Pandey, Sean Fanello, Hyun Soo Park, and Yinda Zhang. One2Avatar: Generative implicit head avatar for few-shot user adaptation. *CoRR*, abs/2402.11909, 2024. 2, 3
- [72] Bowen Zhang, Yiji Cheng, Chunyu Wang, Ting Zhang, Jiaolong Yang, Yansong Tang, Feng Zhao, Dong Chen, and Bainig Guo. Rodinh: High-fidelity 3d avatar generation with diffusion models. *arXiv preprint arXiv:2407.06938*, 2024. 2
- [73] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 8
- [74] Xiaochen Zhao, Jingxiang Sun, Lizhen Wang, Jinli Suo, and Yebin Liu. Invertavatar: Incremental gan inversion for generalized head avatars. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 2, 6, 7, 8
- [75] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, Guidong Wang, and Xu Lan. HeadGAP: Few-shot 3D head avatar via generalizable gaussian priors. *arXiv preprint arXiv:2408.06019*, 2024. 2, 3, 4
- [76] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 5
- [77] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. PointAvatar: Deformable point-based head avatars from videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21057–21067, 2023. 3



- [78] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. MoFaNeRF: Morphable facial neural radiance field. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 6
- [79] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4574–4584, 2022. 2, 3, 5, 6, 7
- [80] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)*, 2022. 5
- [81] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3D gaussian avatars. In *International Conference on 3D Vision (3DV)*, 2025. 4
- [82] Wojciech Zielonka, Timo Bolkart, Thabo Beeler, and Justus Thies. Gaussian eigen models for human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3, 4, 5

## A.6 BROADER IMPACT: ETHICAL CONCERNS

While our research advances the synthesis and capture of realistic 3D avatars, it also raises important ethical considerations. Technologies capable of generating lifelike human representations can be misused for malicious purposes, including misinformation, identity theft, harassment, and other forms of digital deception. Although our work is intended for positive applications such as remote communication and telepresence, we acknowledge the inherent difficulty in preventing misuse.

To mitigate potential harm, we advocate for open and transparent research practices. Sharing our methods and datasets can support the development of safeguards such as digital media forensics and forgery detection systems. At the same time, we recognize that the rapid progress in generative models, particularly diffusion-based approaches, presents significant challenges for digital forensics. As these models continue to improve in generating highly realistic multimodal outputs, including synchronized video and audio, detecting synthetic or manipulated content becomes increasingly difficult.

This remains an urgent and unresolved issue. Current detection tools often struggle to keep pace with generative techniques, and the problem is likely to grow as the quality of generated media improves. We emphasize the need for continued research in both content generation and detection, and we highlight the importance of responsible data usage, ethical oversight, and collaboration across disciplines to ensure that these technologies serve the public good while minimizing potential harm.

## BIBLIOGRAPHY

---

- [1] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y Ogras, and Linjie Luo. “Panohead: Geometry-aware 3d full-head synthesis in 360deg.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 20950–20959 (cit. on pp. 7, 14, 18).
- [2] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. “SINC: Spatial Composition of 3D Human Motions for Simultaneous Action Generation.” In: *International Conference on Computer Vision (ICCV)*. 2023 (cit. on p. 41).
- [3] ShahRukh Athar, Shunsuke Saito, Zhengyu Yang, Stanislav Pidhorskyi, and Chen Cao. “Bridging the Gap: Studio-Like Avatar Creation from a Monocular Phone Capture.” In: *European Conference on Computer Vision (ECCV)*. Springer. 2024, pp. 72–88 (cit. on p. 18).
- [4] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. “Driving-signal aware full-body avatars.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 40.4 (2021), pp. 1–17 (cit. on pp. 15, 17, 28, 29, 41).
- [5] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. “Lumiere: A space-time diffusion model for video generation.” In: *SIGGRAPH Asia 2024 Conference Papers*. 2024, pp. 1–11 (cit. on p. 40).
- [6] Anil Bas and William AP Smith. “What does 2D geometric information really tell us about 3D face shape?” In: *International Journal of Computer Vision* 127 (2019), pp. 1455–1473 (cit. on p. 21).
- [7] Curzio Basso and Alessandro Verri. “Fitting 3D morphable models using implicit representations.” In: *JVRB-Journal of Virtual Reality and Broadcasting* 4.18 (2008) (cit. on p. 8).
- [8] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. “High-quality single-shot capture of facial geometry.” In: *ACM SIGGRAPH 2010 papers*. 2010, pp. 1–9 (cit. on p. 2).
- [9] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul A Beardsley, Craig Gotsman, Robert W Sumner, and Markus H Gross. “High-quality passive facial performance capture using anchor frames.” In: *ACM Trans. Graph.* 30.4 (2011), p. 75 (cit. on p. 2).

- [10] Léore Bensabath, Mathis Petrovich, and Gül Varol. “TMR++: A Cross-Dataset Study for Text-based 3D Human Motion Retrieval.” In: *CVPR Workshop on Human Motion Generation*. 2024 (cit. on p. 41).
- [11] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. “Reanimating faces in images and video.” In: vol. 22. 2003, pp. 641–650 (cit. on pp. 3, 9).
- [12] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. “Exchanging Faces in Images.” In: *Computer Graphics Forum* 23.3 (2004), pp. 669–676 (cit. on pp. 3, 9).
- [13] Volker Blanz and Thomas Vetter. “A morphable model for the synthesis of 3D faces.” In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023, pp. 157–164 (cit. on pp. 3, 7, 8, 10, 11, 13, 21, 23, 31, 32, 34).
- [14] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. “Stable video diffusion: Scaling latent video diffusion models to large datasets.” In: *arXiv preprint arXiv:2311.15127* (2023) (cit. on p. 40).
- [15] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. “Align your latents: High-resolution video synthesis with latent diffusion models.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 22563–22575 (cit. on p. 19).
- [16] Marcel C Buehler, Gengyan Li, Erroll Wood, Leonhard Helming, Xu Chen, Tanmay Shah, Daoye Wang, Stephan Garbin, Sergio Orts-Escolano, Otmar Hilliges, et al. “Cafca: High-quality Novel View Synthesis of Expressive Faces from Casual Few-shot Captures.” In: *SIGGRAPH Conference Papers (SA)*. 2024, pp. 1–12 (cit. on p. 35).
- [17] Ang Cao and Justin Johnson. “HexPlane: A Fast Representation for Dynamic Scenes.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023) (cit. on p. 17).
- [18] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. “Authentic volumetric avatars from a phone scan.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 41.4 (2022), pp. 1–19 (cit. on pp. 18, 19).
- [19] Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. “3D shape regression for real-time facial animation.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 32.4 (2013), pp. 1–10 (cit. on p. 7).

- [20] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "Real-time multi-person 2d pose estimation using part affinity fields." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 7291–7299 (cit. on pp. 7, 29).
- [21] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. "Efficient geometry-aware 3d generative adversarial networks." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 16123–16133 (cit. on pp. 14, 18).
- [22] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. "pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 5799–5809 (cit. on p. 18).
- [23] Aggelina Chatziagapi, Louis-Philippe Morency, Hongyu Gong, Michael Zollhoefer, Dimitris Samaras, and Alexander Richard. "AV-Flow: Transforming Text to Audio-Visual Human-like Interactions." In: *arXiv preprint arXiv:2502.13133* (2025) (cit. on p. 7).
- [24] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. "Photo-Realistic Facial Details Synthesis from Single Image." In: *ICCV*. 2019, pp. 9429–9439 (cit. on p. 12).
- [25] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. "Decision transformer: Reinforcement learning via sequence modeling." In: *Advances in neural information processing systems* 34 (2021), pp. 15084–15097 (cit. on p. 4).
- [26] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. "Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 11594–11604 (cit. on p. 14).
- [27] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. "Monogaussianavatar: Monocular gaussian point-based head avatar." In: *SIGGRAPH Conference Papers (SA)*. 2024, pp. 1–9 (cit. on p. 34).
- [28] Xuangeng Chu and Tatsuya Harada. "Generalizable and animatable gaussian head avatar." In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 57642–57670 (cit. on pp. 7, 34).

- [29] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. “3D U-Net: learning dense volumetric segmentation from sparse annotation.” In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II* 19. Springer. 2016, pp. 424–432 (cit. on p. 19).
- [30] Michael De Smet and Luc Van Gool. “Optimal regions for linear model-based 3D face reconstruction.” In: *Asian conference on computer vision*. Springer. 2010, pp. 276–289 (cit. on p. 8).
- [31] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. “Arcface: Additive angular margin loss for deep face recognition.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4690–4699 (cit. on pp. 2, 22, 38).
- [32] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. “Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 7119–7130 (cit. on pp. 34–36).
- [33] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. “Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set.” In: *Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*. 2019, pp. 0–0 (cit. on pp. 2, 7, 11, 12, 22, 23).
- [34] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” In: *International Conference on Learning Representations (ICLR)*. 2021 (cit. on p. 20).
- [35] Pengfei Dou, Shishir K. Shah, and Ioannis A. Kakadiaris. “End-to-End 3D Face Reconstruction with Deep Neural Networks.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5908–5917. DOI: [10.1109/CVPR.2017.627](https://doi.org/10.1109/CVPR.2017.627) (cit. on p. 12).
- [36] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. “3d morphable face models—past, present, and future.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 39.5 (2020), pp. 1–38 (cit. on pp. 9, 31).
- [37] Patrick Esser, Robin Rombach, and Bjorn Ommer. “Taming transformers for high-resolution image synthesis.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 12873–12883 (cit. on p. 19).



- [38] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. “Fast Dynamic Radiance Fields with Time-Aware Neural Voxels.” In: *SIGGRAPH Asia Conference Papers (SA)*. 2022 (cit. on p. 17).
- [39] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. “Learning an animatable detailed 3D face model from in-the-wild images.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 40.4 (2021), pp. 1–13 (cit. on pp. 2, 7, 11, 12, 15, 22, 23, 32, 35).
- [40] Yao Feng, Jinlong Yang, Timo Bolkart, and Michael J. Black. “DELTA: Deep Learning of Temporally-coherent Animatable Avatars.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023) (cit. on p. 15).
- [41] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. “Capturing and Animation of Body and Clothing from Monocular Video.” In: *Transactions on Graphics, (Proc. SIGGRAPH Asia)* (2022) (cit. on p. 15).
- [42] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter J. B. Hancock, Xiaojun Wu, Qijun Zhao, Paul Koppen, and Matthias Rätzsch. “Evaluation of Dense 3D Reconstruction from 2D Face Images in the Wild.” In: *International Conference on Automatic Face & Gesture Recognition (FG)*. 2018, pp. 780–786 (cit. on pp. 22, 23).
- [43] Richard P. Feynman. *The Feynman Lectures on Physics*. Vol. I, Preface. Addison-Wesley, 1964 (cit. on p. 1).
- [44] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. “Dynamic neural radiance fields for monocular 4d facial avatar reconstruction.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 8649–8658 (cit. on pp. 7, 13, 15, 25, 26, 34, 35).
- [45] Ruiqi Gao\*, Aleksander Holynski\*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole\*. “CAT3D: Create Anything in 3D with Multi-View Diffusion Models.” In: *Advances in Neural Information Processing Systems* (2024) (cit. on p. 20).
- [46] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. “Reconstructing personalized semantic facial nerf models from monocular video.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 41.6 (2022), pp. 1–12 (cit. on p. 13).
- [47] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormaehlen, Patrick Perez, and Christian Theobalt. “Automatic Face Reenactment.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 4217–4224 (cit. on p. 9).

- [48] Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Perez, and Christian Theobalt. “VDub - Modifying Face Video of Actors for Plausible Visual Alignment to a Dubbed Audio Track.” In: 2015, pp. 193–204 (cit. on pp. 9, 17).
- [49] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos P. Zafeiriou. “Fast-GANFIT: Generative Adversarial Network for High Fidelity 3D Face Reconstruction.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021) (cit. on p. 13).
- [50] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. “GANFIT: Generative Adversarial Network Fitting for High Fidelity 3D Face Reconstruction.” In: *CVPR*. 2019 (cit. on p. 13).
- [51] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. “Unsupervised training for 3d morphable model regression.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 8377–8386 (cit. on p. 12).
- [52] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. “Learning neural parametric head models.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 21003–21012 (cit. on p. 9).
- [53] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. “Monophm: Dynamic head reconstruction from monocular videos.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 10747–10758 (cit. on p. 32).
- [54] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. “NPGA: Neural Parametric Gaussian Avatars.” In: *SIGGRAPH Asia Conference Papers (SA)*. 2024. ISBN: 979-8-4007-1131-2/24/12. DOI: [10.1145/3680528.3687689](https://doi.org/10.1145/3680528.3687689) (cit. on pp. 2–4, 7, 10).
- [55] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. “Pixel3DMM: Versatile Screen-Space Priors for Single-Image 3D Face Reconstruction.” In: (2025). URL: <https://arxiv.org/abs/2505.00615> (cit. on pp. 2, 20).
- [56] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative Adversarial Nets.” In: *Advances in Neural Information Processing Systems*. Vol. 27. 2014, pp. 2672–2680 (cit. on p. 18).

- [57] Shrisudhan Govindarajan, Daniel Rebain, Kwang Moo Yi, and Andrea Tagliasacchi. “Radiant Foam: Real-Time Differentiable Ray Tracing.” In: *arXiv preprint arXiv:2502.01157* (2025) (cit. on p. 38).
- [58] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. “Neural head avatars from monocular rgb videos.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 18653–18664 (cit. on pp. 7, 10, 13, 25, 26, 35).
- [59] Zekai Gu et al. “Diffusion as Shader: 3D-aware Video Diffusion for Versatile Video Generation Control.” In: *arXiv preprint arXiv:2501.03847* (2025) (cit. on p. 40).
- [60] Minghao Guo, Bohan Wang, Kaiming He, and Wojciech Matusik. “Tetsphere splatting: Representing high-quality geometry with lagrangian volumetric meshes.” In: *International Conference on Learning Representations (ICLR)* (2025) (cit. on p. 38).
- [61] Pengsheng Guo and Alexander G. Schwing. “Variational Rectified Flow Matching.” In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2025 (cit. on p. 17).
- [62] Vladimir Guzov, Yifeng Jiang, Fangzhou Hong, Gerard Pons-Moll, Richard Newcombe, C Karen Liu, Yuting Ye, and Lingni Ma. “HMD<sup>2</sup>: Environment-aware Motion Generation from Single Egocentric Head-Mounted Device.” In: *International Conference on 3D Vision (3DV)* (2025) (cit. on p. 20).
- [63] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising Diffusion Probabilistic Models.” In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851 (cit. on pp. 17, 20).
- [64] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. “Video diffusion models.” In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 8633–8646 (cit. on pp. 19, 40).
- [65] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. “Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 634–644 (cit. on pp. 15, 28, 29).
- [66] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. “Avatar Digitization from a Single Image for Real-Time Rendering.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 36.6 (2017), 195:1–195:14 (cit. on p. 10).

- [67] Jin Huang, Xiaohan Shi, Xinguo Liu, Kun Zhou, Li-Yi Wei, Shang-Hua Teng, Hujun Bao, Baining Guo, and Harry Shum. "Subspace gradient domain mesh deformation." In: *Transactions on Graphics, (Proc. SIGGRAPH)* (2006) (cit. on p. 15).
- [68] Alexandru-Eugen Ichim, Petr Kadleček, Ladislav Kavan, and Mark Pauly. "Phace: Physics-based face modeling and animation." In: *ACM Transactions on Graphics (ToG)* 36.4 (2017), pp. 1–14 (cit. on p. 9).
- [69] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. "Humanrf: High-fidelity neural radiance fields for humans in motion." In: *Transactions on Graphics, (Proc. SIGGRAPH)* 42.4 (2023), pp. 1–12 (cit. on pp. 17, 28).
- [70] Alec Jacobson, Ilya Baran, Jovan Popović, and Olga Sorkine-Hornung. "Bounded biharmonic weights for real-time deformation." In: *Transactions on Graphics, (Proc. SIGGRAPH)* (2011) (cit. on p. 15).
- [71] Ying Jiang, Chang Yu, Tianyi Xie, Xuan Li, Yutao Feng, Huamin Wang, Minchen Li, Henry Lau, Feng Gao, Yin Yang, et al. "Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality." In: *SIGGRAPH Conference Papers (SA)*. 2024, pp. 1–1 (cit. on p. 16).
- [72] Tao Ju, Scott Schaefer, and Joe D. Warren. "Mean value coordinates for closed triangular meshes." In: *Transactions on Graphics, (Proc. SIGGRAPH)* (2005) (cit. on p. 15).
- [73] Berna Kabadayi, Wojciech Zielonka, Bharat Lal Bhatnagar, Gerard Pons-Moll, and Justus Thies. "Gan-avatar: Controllable personalized gan-based human head avatar." In: *International Conference on 3D Vision (3DV)*. IEEE. 2024, pp. 882–892 (cit. on pp. 7, 14, 18).
- [74] James T. Kajiya. "The Rendering Equation." In: *Transactions on Graphics, (Proc. SIGGRAPH)* 20 (1986), pp. 143–150 (cit. on p. 15).
- [75] Yash Kant, Ethan Weber, Jin Kyu Kim, Rawal Khirodkar, Su Zhaoen, Julieta Martinez, Igor Gilitschenski, Shunsuke Saito, and Timur Bagautdinov. "Pippo: High-Resolution Multi-View Humans from a Single Image." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025 (cit. on pp. 7, 20).
- [76] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Alias-free generative adversarial networks." In: *Advances in neural information processing systems* 34 (2021), pp. 852–863 (cit. on p. 17).

- [77] Tero Karras, Samuli Laine, and Timo Aila. “A style-based generator architecture for generative adversarial networks.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4401–4410 (cit. on pp. 2, 17, 34).
- [78] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. “Analyzing and improving the image quality of stylegan.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 8110–8119 (cit. on pp. 14, 17).
- [79] Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. “GMD: Controllable Human Motion Synthesis via Guided Diffusion Models.” In: (2024) (cit. on p. 20).
- [80] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. “3d gaussian splatting for real-time radiance field rendering.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 42.4 (2023), pp. 139–1 (cit. on pp. 7, 13–15, 17, 31, 34, 35).
- [81] Hyeonwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. “Deep video portraits.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 37.4 (2018), pp. 1–14 (cit. on pp. 9, 10).
- [82] Diederik P Kingma and Max Welling. “Auto-Encoding Variational Bayes.” In: *International Conference on Learning Representations (ICLR)*. 2014 (cit. on pp. 17, 18).
- [83] Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. “DiffusionAvatars: Deferred Diffusion for High-fidelity 3D Head Avatars.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024 (cit. on pp. 7, 20).
- [84] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. “GGHead: Fast and Generalizable 3D Gaussian Heads.” In: *SIGGRAPH Asia Conference Papers (SA)*. SA '24. New York, NY, USA: Association for Computing Machinery, 2024. ISBN: 9798400711312 (cit. on pp. 7, 20).
- [85] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. “NeRSemble: Multi-View Radiance Field Reconstruction of Human Heads.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 42.4 (2023). ISSN: 0730-0301 (cit. on pp. 32, 40).
- [86] Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. “Avat3r: Large Animatable Gaussian Reconstruction Model for High-fidelity 3D

- Head Avatars." In: (2025). arXiv: 2502.20220 [cs.CV]. URL: <https://arxiv.org/abs/2502.20220> (cit. on pp. 3, 20).
- [87] Paul Koppen, Zhen-Hua Feng, Josef Kittler, Muhammad Awais, William Christmas, Xiao-Jun Wu, and He-Feng Yin. "Gaussian mixture 3D morphable face model." In: *Pattern Recognition* 74 (2018), pp. 617–628 (cit. on p. 9).
- [88] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. "Modular Primitives for High-Performance Differentiable Rendering." In: *Transactions on Graphics, (Proc. SIGGRAPH)* 39.6 (2020) (cit. on p. 11).
- [89] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. "Building machines that learn and think like people." In: *Behavioral and Brain Sciences* 40 (2017), e253 (cit. on p. 41).
- [90] Christoph Lassner and Michael Zollhofer. "Pulsar: Efficient sphere-based neural rendering." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 1440–1449 (cit. on pp. 11, 16).
- [91] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. "AvatarMe: Realistically Renderable 3D Facial Reconstruction In-the-Wild." In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 760–769 (cit. on pp. 13, 18).
- [92] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P. Zafeiriou. "AvatarMe++: Facial Shape and BRDF Inference with Photorealistic Rendering-Aware GANs." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021) (cit. on p. 13).
- [93] Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. "Realtime facial animation with on-the-fly correctives." In: *Transactions on Graphics, (Proc. SIGGRAPH)* 32.4 (2013), pp. 42–1 (cit. on p. 7).
- [94] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. "Tava: Template-free animatable volumetric actors." In: *European Conference on Computer Vision*. Springer. 2022, pp. 419–436 (cit. on pp. 15, 28).
- [95] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. "Learning a model of facial shape and expression from 4D scans." In: *Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36.6 (2017), pp. 194–1 (cit. on pp. 2, 3, 5, 7–9, 13, 21–23, 25, 27, 28, 34, 35).



- [96] Tianye Li, Shichen Liu, Timo Bolkart, Jiayi Liu, Hao Li, and Yajie Zhao. "Topologically consistent multi-view face inference using volumetric sampling." In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 3824–3834 (cit. on p. 2).
- [97] Tianye Li et al. "Neural 3D Video Synthesis from Multi-view Video." In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022) (cit. on p. 17).
- [98] Zhe Li, Yipengjing Sun, Zerong Zheng, Lizhen Wang, Shengping Zhang, and Yebin Liu. "Animatable and Relightable Gaussians for High-fidelity Human Avatar Modeling." In: *arXiv preprint arXiv:2311.16096* (2023) (cit. on pp. 15, 16, 28–32, 35, 41).
- [99] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. "Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes." In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 6494–6504 (cit. on p. 15).
- [100] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. "Llm-grounded video diffusion models." In: *arXiv preprint arXiv:2309.17444* (2023) (cit. on p. 40).
- [101] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. "Flow Matching for Generative Modeling." In: *International Conference on Learning Representations (ICLR)*. 2023 (cit. on p. 17).
- [102] Qiang Liu. "Rectified Flow: A Marginal Preserving Approach to Optimal Transport." In: *CoRR abs/2209.14577* (2022) (cit. on pp. 7, 17).
- [103] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. "Soft rasterizer: A differentiable renderer for image-based 3d reasoning." In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 7708–7717 (cit. on p. 11).
- [104] Xingchao Liu, Chengyue Gong, and Qiang Liu. "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow." In: *International Conference on Learning Representations (ICLR)*. 2023 (cit. on p. 17).
- [105] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. "Neural volumes: Learning dynamic renderable volumes from images." In: *Transactions on Graphics, (Proc. SIGGRAPH)* (2019) (cit. on pp. 17, 19).
- [106] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. "Mixture of Volumetric Primitives for Efficient Neural Rendering." In: *Transactions on Graphics, (Proc. SIGGRAPH)* 40.4 (2021) (cit. on pp. 3, 4, 7, 15, 17, 19, 28–30).

- [107] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. "SMPL: A Skinned Multi-Person Linear Model." In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2* (2015) (cit. on pp. 15, 16).
- [108] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. "SMPL: A Skinned Multi-Person Linear Model." In: *Transactions on Graphics, (Proc. SIGGRAPH Asia)* 34.6 (Oct. 2015), 248:1–248:16 (cit. on pp. 2, 3).
- [109] Birgit Lugin, Catherine Pelachaud, and David Traum. *The handbook on socially interactive agents: 20 years of research on embodied conversational agents, intelligent virtual agents, and social robotics volume 2: interactivity, platforms, application*. ACM, 2022 (cit. on pp. 2, 4, 7).
- [110] Haimin Luo, Ouyang Min, Zijun Zhao, Suyi Jiang, Longwen Zhang, Qixuan Zhang, Wei Yang, Lan Xu, and Jingyi Yu. "GaussianHair: Hair Modeling and Rendering with Light-aware Gaussians." In: *ArXiv* (2024) (cit. on p. 16).
- [111] Weijie Lyu, Yi Zhou, Ming-Hsuan Yang, and Zhixin Shu. "FaceLift: Single Image to 3D Head with View Generation and GS-LRM." In: *arXiv preprint arXiv:2412.17812* (2024) (cit. on p. 20).
- [112] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. "The Power of Points for Modeling Humans in Clothing." In: *International Conference on Computer Vision (ICCV)* (2021), pp. 10954–10964 (cit. on pp. 15, 16).
- [113] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J Black. "The power of points for modeling humans in clothing." In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 10974–10984 (cit. on p. 3).
- [114] Shugao Ma, Tomas Simon, Jason M. Saragih, Dawei Wang, Yuecheng Li, Fernando De la Torre, and Yaser Sheikh. "Pixel Codec Avatars." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, 2021 (cit. on pp. 11, 17, 19, 34).
- [115] Julieta Martinez et al. "Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars." In: *NeurIPS Track on Datasets and Benchmarks* (2024) (cit. on pp. 17, 34, 40).
- [116] Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. "KeypointNeRF: Generalizing Image-based Volumetric Avatars using Relative Spatial Encoding of Keypoints." In: (2022) (cit. on p. 15).

- [117] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis.” In: *European Conference on Computer Vision (ECCV)*. 2020 (cit. on pp. 7, 13, 15, 25, 27, 32, 34).
- [118] Araceli Morales, Gemma Piella, and Federico M. Sukno. *Survey on 3D face reconstruction from uncalibrated images*. 2021. arXiv: 2011.05740 [cs.CV] (cit. on p. 9).
- [119] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. “Instant neural graphics primitives with a multiresolution hash encoding.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 41.4 (2022), pp. 1–15 (cit. on pp. 3, 25, 27).
- [120] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. “pa-GAN: Real-Time Avatars Using Dynamic Textures.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 37.6 (2018), 258:1–258:12 (cit. on p. 10).
- [121] Thomas Neumann, Kiran Varanasi, Stephan Wenger, Markus Wacker, Marcus Magnor, and Christian Theobalt. “Sparse Localized Deformation Components.” In: *Transactions on Graphics, (Proc. SIGGRAPH Asia)* 32.6 (Nov. 2013) (cit. on pp. 8, 33).
- [122] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. “From audio to photoreal embodiment: Synthesizing humans in conversations.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 1001–1010 (cit. on pp. 2, 4, 5, 7, 41).
- [123] Alexander Quinn Nichol and Prafulla Dhariwal. “Improved Denoising Diffusion Probabilistic Models.” In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8162–8171 (cit. on pp. 17, 20).
- [124] Jesús Rodríguez Nieto and Antonio Susín. “Deformation Models: Tracking, Animation and Applications.” In: *Deformation Models: Tracking, Animation and Applications*. Springer, 2012, pp. 75–99 (cit. on p. 16).
- [125] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. “Npms: Neural parametric models for 3d deformable shapes.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 12695–12705 (cit. on p. 9).

- [126] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. “Ash: Animatable gaussian splats for efficient and photoreal human rendering.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 1165–1175 (cit. on pp. 7, 16, 31).
- [127] Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, et al. “Stabilizing transformers for reinforcement learning.” In: *International conference on machine learning*. PMLR. 2020, pp. 7487–7498 (cit. on p. 4).
- [128] Keunhong Park, U. Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. “Nerfies: Deformable Neural Radiance Fields.” In: *International Conference on Computer Vision (ICCV)* (2020), pp. 5845–5854 (cit. on p. 15).
- [129] Keunhong Park, U. Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. “HyperNeRF.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 40 (2021), pp. 1–12 (cit. on p. 15).
- [130] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000 (cit. on p. 41).
- [131] William Peebles and Saining Xie. “Scalable Diffusion Models with Transformers.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023. DOI: [10.1109/ICCV51070.2023.00387](https://doi.org/10.1109/ICCV51070.2023.00387) (cit. on p. 20).
- [132] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. “Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies.” In: *ICCV*. 2021 (cit. on p. 15).
- [133] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. “Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 9050–9059 (cit. on p. 15).
- [134] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. “Amp: Adversarial motion priors for stylized physics-based character control.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 40.4 (2021), pp. 1–20 (cit. on p. 4).
- [135] Yicong Peng, Yichao Yan, Shengqi Liu, Yuhao Cheng, Shanyan Guan, Bowen Pan, Guangtao Zhai, and Xiaokang Yang. “Cagenerf: Cage-based neural radiance field for generalized 3d

- deformation and animation." In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 31402–31415 (cit. on p. 17).
- [136] Pentland, Moghaddam, and Starner. "View-based and modular eigenspaces for face recognition." In: *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 1994, pp. 84–91 (cit. on p. 7).
- [137] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. "Film: Visual reasoning with a general conditioning layer." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018 (cit. on p. 18).
- [138] Mathis Petrovich, Michael J. Black, and Gül Varol. "TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis." In: *International Conference on Computer Vision (ICCV)*. 2023 (cit. on p. 41).
- [139] Malte Prinzler, Otmar Hilliges, and Justus Thies. "Diner: Depth-aware image-based neural radiance fields." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 12449–12459 (cit. on pp. 14, 15).
- [140] Sergey Prokudin, Qianli Ma, Maxime Raafat, Julien Valentin, and Siyu Tang. "Dynamic Point Fields." In: *International Conference on Computer Vision (ICCV)*. 2023, pp. 7964–7976 (cit. on p. 16).
- [141] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. "Pointnet++: Deep hierarchical feature learning on point sets in a metric space." In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 35).
- [142] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. "Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 20299–20309 (cit. on pp. 4, 10, 15, 16, 31, 32, 34).
- [143] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. "3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 5020–5030 (cit. on pp. 28, 29).
- [144] Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, et al. "Drivable volumetric avatars using texel-aligned features." In: *ACM SIGGRAPH 2022 conference proceedings*. 2022, pp. 1–9 (cit. on pp. 17, 28, 41).

- [145] Elad Richardson, Matan Sela, and Ron Kimmel. “3D Face Reconstruction by Learning from Synthetic Data.” In: *International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 460–469. DOI: [10.1109/3DV.2016.54](https://doi.org/10.1109/3DV.2016.54) (cit. on p. 12).
- [146] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. “Learning Detailed Face Reconstruction from a Single Image.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5553–5562. DOI: [10.1109/CVPR.2017.589](https://doi.org/10.1109/CVPR.2017.589) (cit. on p. 12).
- [147] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. “Pivotal tuning for latent-based editing of real images.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 42.1 (2022), pp. 1–13 (cit. on p. 35).
- [148] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. “High-Resolution Image Synthesis with Latent Diffusion Models.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695 (cit. on pp. 7, 17, 19, 20).
- [149] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019 (cit. on p. 41).
- [150] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. “Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization.” In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 2304–2314 (cit. on p. 3).
- [151] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. “Relightable gaussian codec avatars.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 130–141 (cit. on pp. 3, 7, 11, 16, 17, 19, 31, 35).
- [152] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. “Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 84–93 (cit. on p. 3).
- [153] Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. “Photorealistic facial texture inference using deep neural networks.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5144–5153 (cit. on p. 13).
- [154] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. “SCANimate: Weakly supervised learning of skinned clothed avatar networks.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 2886–2897 (cit. on p. 3).



- [155] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. “Learning to regress 3D face shape and expression from an image without 3D supervision.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 7763–7772 (cit. on pp. [12](#), [22](#)).
- [156] Jean-Paul Sartre. *Existentialism is a Humanism*. Trans. by Carol Macomber. New Haven: Yale University Press, 2007 (cit. on p. [1](#)).
- [157] Jack Saunders, Charlie Hewitt, Yanan Jian, Marek Kowalski, Tadas Baltrusaitis, Yiye Chen, Darren Cosker, Virginia Estellers, Nicholas Gyde, Vinay P Namboodiri, et al. “GASP: Gaussian Avatars with Synthetic Priors.” In: *arXiv preprint arXiv:2412.07739* (2024) (cit. on pp. [2](#), [3](#)).
- [158] Bernhard Schölkopf. *Causality for Machine Learning*. Preliminary version, MIT Press (forthcoming). 2022 (cit. on p. [41](#)).
- [159] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A Unified Embedding for Face Recognition and Clustering.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 815–823 (cit. on pp. [12](#), [13](#)).
- [160] Christoph Schuhmann et al. *LAION-5B: An open large-scale dataset for training next generation image-text models*. 2022. arXiv: [2210.08402 \[cs.CV\]](#) (cit. on p. [40](#)).
- [161] Matan Sela, Elad Richardson, and Ron Kimmel. “Unrestricted facial geometry reconstruction using image-to-image translation.” In: *International Conference on Computer Vision (ICCV)*. 2017, pp. 1576–1585 (cit. on p. [35](#)).
- [162] Mike Seymour, Chris Evans, and Kim Libreri. “Meet mike: epic avatars.” In: *ACM SIGGRAPH 2017 VR Village*. 2017, pp. 1–2 (cit. on p. [34](#)).
- [163] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. “Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 1606–1616 (cit. on pp. [4](#), [35](#)).
- [164] Vanessa Sklyarova, Jenya Chelishev, Andreea Dogaru, Igor Medvedev, Victor Lempitsky, and Egor Zakharov. “Neural hair-cut: Prior-guided strand-based hair reconstruction.” In: *International Conference on Computer Vision (ICCV)*. 2023, pp. 19762–19773 (cit. on p. [7](#)).

- [165] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. “Deep Unsupervised Learning using Nonequilibrium Thermodynamics.” In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. JMLR Proceedings. 2015, pp. 2256–2265 (cit. on pp. 17, 20).
- [166] Jiaming Song, Chenlin Meng, and Stefano Ermon. “Denoising Diffusion Implicit Models.” In: *International Conference on Learning Representations*. OpenReview.net, 2021 (cit. on pp. 17, 20).
- [167] Yang Song and Stefano Ermon. “Generative Modeling by Estimating Gradients of the Data Distribution.” In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019, pp. 11895–11907 (cit. on pp. 17, 20).
- [168] Sebastian Starke, Ian Mason, and Taku Komura. “Deepphase: Periodic autoencoders for learning motion phase manifolds.” In: *ACM Transactions on Graphics (ToG)* 41.4 (2022), pp. 1–13 (cit. on p. 4).
- [169] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. “Neural state machine for character-scene interactions.” In: *ACM Transactions on Graphics* 38.6 (2019), p. 178 (cit. on p. 4).
- [170] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi A Zaman. “Local motion phases for learning multi-contact character movements.” In: *ACM Trans. Graph.* 39.4 (2020), p. 54 (cit. on p. 4).
- [171] Shih-Yang Su, Timur M. Bagautdinov, and Helge Rhodin. “DANBO: Disentangled Articulated Neural Body Representations via Graph Neural Networks.” In: *European Conference on Computer Vision (ECCV)*. 2022 (cit. on p. 15).
- [172] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. “Npc: Neural point characters from video.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 14795–14805 (cit. on pp. 15, 16, 28).
- [173] Shih-Yang Su, Frank Yu, Michael Zollhoefer, and Helge Rhodin. “A-NeRF: Articulated Neural Radiance Fields for Learning Human Shape, Appearance, and Pose.” In: *Neural Information Processing Systems*. 2021 (cit. on p. 15).
- [174] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. “Next3d: Generative neural texture rasterization for 3d-aware head avatars.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 20991–21002 (cit. on pp. 34–36).

- [175] Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B Lindell. "CAP4D: Creating Animatable 4D Portrait Avatars with Morphable Multi-View Diffusion Models." In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025) (cit. on pp. 2–4, 20).
- [176] J Rafael Tena, Fernando De la Torre, and Iain A Matthews. "Interactive region-based linear 3d face models." In: *Transactions on Graphics, (Proc. SIGGRAPH)* 30.4 (2011), p. 76 (cit. on p. 8).
- [177] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. "State of the art on neural rendering." In: *Computer Graphics Forum*. Vol. 39. 2. Wiley Online Library. 2020, pp. 701–727 (cit. on pp. 13, 31).
- [178] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. "Advances in neural rendering." In: *Computer Graphics Forum*. Vol. 41. 2. Wiley Online Library. 2022, pp. 703–735 (cit. on pp. 13, 31).
- [179] Ayush Tewari, Michael Zollhöfer, Hyeonwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. "InverseFaceNet: Deep Monocular Inverse Face Rendering." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 4625–4634. DOI: [10.1109/CVPR.2018.00485](https://doi.org/10.1109/CVPR.2018.00485) (cit. on p. 12).
- [180] *The Holy Bible, English Standard Version*. Genesis 1:26–27. Crossway, 2001 (cit. on p. 1).
- [181] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt. "Real-time Expression Transfer for Facial Reenactment." In: *Transactions on Graphics, (Proc. SIGGRAPH)* 34.6 (2015) (cit. on p. 7).
- [182] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. "Face2Face: Real-time Face Capture and Reenactment of RGB Videos." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 (cit. on pp. 3, 7, 9–13, 23, 32).
- [183] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. "FaceVR: Real-Time Gaze-Aware Facial Reenactment in Virtual Reality." In: *Transactions on Graphics, (Proc. SIGGRAPH)* (2018) (cit. on p. 9).
- [184] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. "Neural Voice Puppetry: Audio-driven Facial Reenactment." In: *European Conference on Computer Vision (ECCV)* (2020) (cit. on pp. 9, 10).

- [185] Justus Thies, Michael Zollhöfer, and Matthias Nießner. “Deferred neural rendering: Image synthesis using neural textures.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 38.4 (2019), pp. 1–12 (cit. on pp. 9, 10).
- [186] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. “Real-time Expression Transfer for Facial Reenactment.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 34.6 (2015) (cit. on pp. 9–11).
- [187] Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Niessner. “HeadOn: Real-time Reenactment of Human Portrait Videos.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 37.4 (2018), 1–13 (cit. on pp. 9, 10).
- [188] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. “Extreme 3D Face Reconstruction: Seeing Through Occlusions.” In: *CVPR*. 2018 (cit. on p. 12).
- [189] Aaron Van Den Oord, Oriol Vinyals, and Koray Kavukcuoglu. “Neural Discrete Representation Learning.” In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017 (cit. on p. 14).
- [190] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is all you need.” In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 4).
- [191] Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. “Styleavatar: Real-time photo-realistic portrait avatar from a single video.” In: *ACM SIGGRAPH 2023 Conference Proceedings*. 2023, pp. 1–10 (cit. on pp. 14, 31, 34).
- [192] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. “Rodin: A generative model for sculpting 3d digital avatars using diffusion.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 4563–4573 (cit. on p. 35).
- [193] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. “Uni-animate: Taming unified video diffusion models for consistent human image animation.” In: *arXiv preprint arXiv:2406.01188* (2024) (cit. on p. 20).
- [194] Yifan Wang, Noam Aigerman, Vladimir G. Kim, Siddhartha Chaudhuri, and Olga Sorkine-Hornung. “Neural Cages for Detail-Preserving 3D Deformations.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 75–84 (cit. on p. 17).

- [195] Yifan Wang, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. “Differentiable Surface Splatting for Point-based Geometry Processing.” In: *SIGGRAPH Asia Conference Papers (SA)*. 2019 (cit. on p. 16).
- [196] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. “Real-time performance-based facial animation.” In: *Transactions on Graphics, (Proc. SIGGRAPH)*. Vol. 30. 4. 2011 (cit. on p. 9).
- [197] Thibaut Weise, Hao Li, Luc J. Van Gool, and Mark Pauly. “Face/Off: live facial puppetry.” In: *SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*. 2009, pp. 7–16 (cit. on p. 9).
- [198] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. “Fake it till you make it: face analysis in the wild using synthetic data alone.” In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 3681–3691 (cit. on pp. 2, 7, 11, 35).
- [199] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. “4D Gaussian Splatting for Real-Time Dynamic Scene Rendering.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024) (cit. on p. 17).
- [200] Cheng-hsin Wu et al. “Multiface: A Dataset for Neural Face Rendering.” In: *arXiv*. 2022. DOI: [10.48550/ARXIV.2207.11243](https://arxiv.org/abs/2207.11243). URL: <https://arxiv.org/abs/2207.11243> (cit. on p. 34).
- [201] Hongchi Xia, Entong Su, Marius Memmel, Arhan Jain, Raymond Yu, Numfor Mbiziwo-Tiapo, Ali Farhadi, Abhishek Gupta, Shenlong Wang, and Wei-Chiu Ma. “DRAWER: Digital Reconstruction and Articulation With Environment Realism.” In: *arXiv preprint arXiv:2504.15278* (2025) (cit. on p. 41).
- [202] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. “Flashavatar: High-fidelity head avatar with efficient gaussian embedding.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 1802–1812 (cit. on pp. 2, 4, 15, 16, 34, 35).
- [203] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. “Physgaussian: Physics-integrated 3d gaussians for generative dynamics.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 4389–4398 (cit. on p. 16).
- [204] Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. “InterTrack: Tracking Human Object Interaction without Object Templates.” In: *International Conference on 3D Vision (3DV)*. 2025 (cit. on p. 41).

- [205] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. "Econ: Explicit clothed humans optimized via normal integration." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 512–523 (cit. on p. 3).
- [206] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. "Icon: Implicit clothed humans obtained from normals." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2022, pp. 13286–13296 (cit. on p. 3).
- [207] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. "Point-NeRF: Point-based Neural Radiance Fields." In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 5428–5438 (cit. on p. 15).
- [208] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. "Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 1931–1941 (cit. on pp. 2, 10, 15, 16, 31).
- [209] Yuelang Xu, Lizhen Wang, Zerong Zheng, Zhaoqi Su, and Yebin Liu. "3d gaussian parametric head model." In: *European Conference on Computer Vision (ECCV)*. Springer. 2024, pp. 129–147 (cit. on pp. 3, 7, 14, 34).
- [210] Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Han Huang, Guojun Qi, and Yebin Liu. "Latentavatar: Learning latent expression code for expressive neural head avatar." In: *SIGGRAPH Conference Papers (SA)*. 2023, pp. 1–10 (cit. on pp. 14, 18).
- [211] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. "High-Fidelity Facial Reflectance and Geometry Inference from an Unconstrained Image." In: *Transactions on Graphics, (Proc. SIGGRAPH)* 37.4 (2018), 162:1–162:14 (cit. on pp. 10, 13).
- [212] Chenglin Yang, Yinghao Xu, J. Yu, Yebin Liu, and J. Yu. "Humans as Points: Submanifold NeRF for Clothed Human Reconstruction." In: *CVPR*. 2022 (cit. on p. 15).
- [213] Xiuyu Yang, Yunze Man, Junkun Chen, and Yu-Xiong Wang. "SceneCraft: Layout-guided 3D scene generation." In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 82060–82084 (cit. on p. 41).
- [214] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. "Real-time Photorealistic Dynamic Scene Representation and Rendering with 4D Gaussian Splatting." In: *International Conference on Learning Representations (ICLR)*. 2024 (cit. on p. 17).



- [215] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. “Multiview neural surface reconstruction by disentangling geometry and appearance.” In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 2492–2502 (cit. on p. 14).
- [216] Egor Zakharov, Vanessa Sklyarova, Michael Black, Giljoo Nam, Justus Thies, and Otmar Hilliges. “Human hair reconstruction with strand-aligned 3d gaussians.” In: *European Conference on Computer Vision (ECCV)*. Springer. 2024, pp. 409–425 (cit. on p. 7).
- [217] Bowen Zhang, Yiji Cheng, Chunyu Wang, Ting Zhang, Jiaolong Yang, Yansong Tang, Feng Zhao, Dong Chen, and Baining Guo. “Rodinhd: High-fidelity 3d avatar generation with diffusion models.” In: *European Conference on Computer Vision (ECCV)*. Springer. 2024, pp. 465–483 (cit. on p. 35).
- [218] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. “Text2nerf: Text-driven 3d scene generation with neural radiance fields.” In: *IEEE Transactions on Visualization and Computer Graphics* 30.12 (2024), pp. 7749–7762 (cit. on p. 15).
- [219] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. “GS-LRM: Large Reconstruction Model for 3D Gaussian Splatting.” In: *European Conference on Computer Vision* (2024) (cit. on p. 20).
- [220] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018 (cit. on p. 29).
- [221] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya A. Petrov, Vladimir Guзов, Helisa Dharmo, Eduardo Pérez Pellitero, and Gerard Pons-Moll. “FORCE: Dataset and Method for Intuitive Physics Guided Human-object Interaction.” In: *International Conference on 3D Vision (3DV)*. 2025 (cit. on p. 41).
- [222] Xiaochen Zhao, Jingxiang Sun, Lizhen Wang, Jinli Suo, and Yebin Liu. “InvertAvatar: Incremental GAN Inversion for Generalized Head Avatars.” In: *SIGGRAPH Conference Papers (SA)*. 2024, pp. 1–10 (cit. on pp. 7, 34–36).
- [223] Xiaochen Zhao, Lizhen Wang, Jingxiang Sun, Hongwen Zhang, Jinli Suo, and Yebin Liu. “Havatar: High-fidelity head avatar via facial model conditioned neural radiance field.” In: *Transactions on Graphics, (Proc. SIGGRAPH)* 43.1 (2023), pp. 1–16 (cit. on pp. 10, 14).

- [224] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. "Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 19680–19690 (cit. on pp. 7, 16, 31).
- [225] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, et al. "Headgap: Few-shot 3d head avatar via generalizable gaussian priors." In: 2025 (cit. on pp. 3, 7, 14, 34).
- [226] Yang Zheng, Qingqing Zhao, Guandao Yang, Wang Yifan, Donglai Xiang, Florian Dubost, Dmitry Lagun, Thabo Beeler, Federico Tombari, Leonidas Guibas, et al. "Physavatar: Learning the physics of dressed 3d avatars from visual observations." In: *European Conference on Computer Vision (ECCV)*. Springer. 2024, pp. 262–284 (cit. on p. 16).
- [227] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. "Im avatar: Implicit morphable head avatars from videos." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 13545–13555 (cit. on pp. 3, 7, 14, 25, 26, 35).
- [228] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. "Pointavatar: Deformable point-based head avatars from videos." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 21057–21067 (cit. on pp. 14–16).
- [229] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. "Structured local radiance fields for human avatar modeling." In: (2022), pp. 15893–15903 (cit. on p. 15).
- [230] Hao Zhu, Haotian Yang, Longwei Guo, Yidi Zhang, Yanru Wang, Mingkai Huang, Menghua Wu, Qiu Shen, Ruigang Yang, and Xun Cao. "FaceScape: 3D Facial Dataset and Benchmark for Single-View 3D Face Reconstruction." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2023) (cit. on p. 40).
- [231] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. "Drivable 3d gaussian avatars." In: *International Conference on 3D Vision (3DV)* (2025) (cit. on pp. 1, 5, 7, 15, 31, 38, 41).
- [232] Wojciech Zielonka, Timo Bolkart, Thabo Beeler, and Justus Thies. "Gaussian eigen models for human heads." In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025) (cit. on pp. 1–4, 7, 16, 34–36, 38).

- [233] Wojciech Zielonka, Timo Bolkart, and Justus Thies. "Towards Metrical Reconstruction of Human Faces." In: *European Conference on Computer Vision (ECCV)*. Computer Vision Foundation / IEEE, 2022 (cit. on pp. [1](#), [2](#), [4](#), [7](#), [25](#), [32](#), [35](#), [37](#)).
- [234] Wojciech Zielonka, Timo Bolkart, and Justus Thies. "Instant volumetric head avatars." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 4574–4584 (cit. on pp. [1–5](#), [7](#), [10](#), [14](#), [15](#), [32](#), [34](#), [35](#), [38](#)).
- [235] Wojciech Zielonka, Stephan J Garbin, Alexandros Lattas, George Kopanas, Paulo Gotardo, Thabo Beeler, Justus Thies, and Timo Bolkart. "Synthetic Prior for Few-Shot Drivable Head Avatar Inversion." In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025) (cit. on pp. [1–5](#), [7](#), [14](#), [16](#), [19](#), [33](#), [39](#)).
- [236] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. "State of the art on monocular 3D face reconstruction, tracking, and applications." In: *Computer graphics forum*. Vol. 37. 2. Wiley Online Library. 2018, pp. 523–550 (cit. on pp. [7](#), [9](#), [10](#), [31](#)).