How to Build Digital Humans?
From Priors to Photorealistic Avatars

Wojciech Zielonka[1][†], Tobias Kirschstein[2][†], Timo Bolkart[3][†], Simon Giebenhain[2], Vanessa Sklyarova[5,8], Xiang Deng[4],
Donglai Xiang[6], Shunsuke Saito[1], Yebin Liu[4], Matthias Niessner[2], Justus Thies[7][†]

[1]Meta   [2]Technical University of Munich   [3]Google   [4]Tsinghua University
[5]Max Planck Institute for Intelligent Systems   [6]NVIDIA   [7]Technical University of Darmstadt   [8]ETH Zurich

[†] These authors contributed equally.

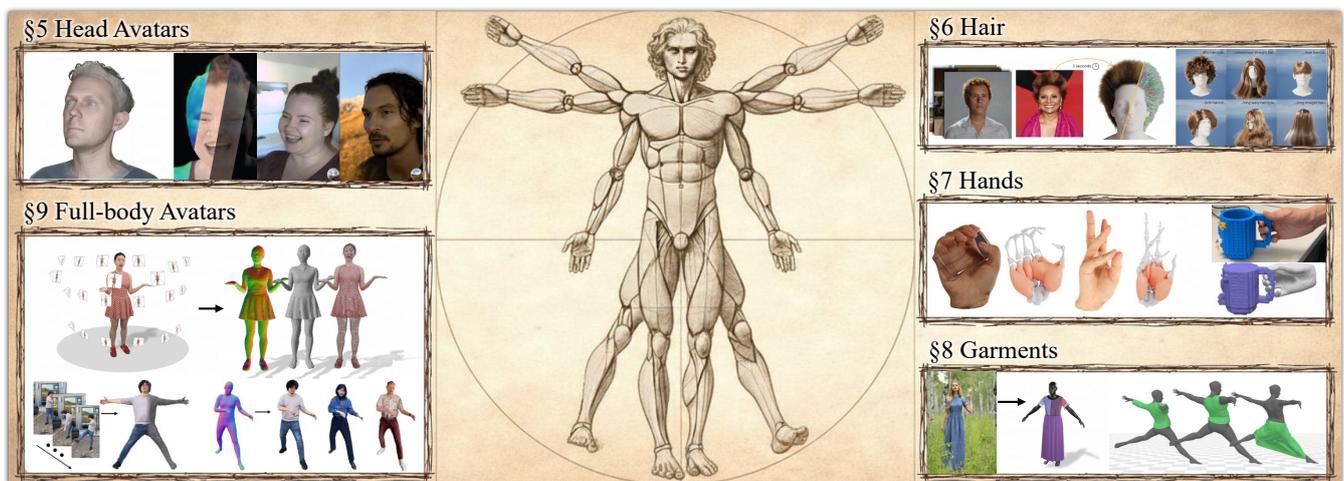https://zielon.github.io/how-to-build-digital-humans/

Figure 1: *Building digital avatars requires considering many components, such as the **human head, hair, hands, and garments**, and ultimately **full-body avatars**. Much like da Vinci's Vitruvian Man symbolizes the ideal proportions and unity of the human form, the construction of digital avatars demands a holistic approach in which each part integrates into a coherent whole. Each of these elements presents distinct challenges and numerous potential design choices across different stages. In this report, we introduce and examine the details of each creation stage, develop a practical taxonomy of the current literature, and discuss future directions in the field of digital avatars. The background image was generated using Google Gemini.* [QKS*24, SSS*24, CHW*25, GLK*25, RWF*25, SZH*23, LZQ*22, FPK*24, BXX*25, GBH23].

Abstract
This state-of-the-art report provides an overview of controllable 3D human avatar creation. We describe current 3D avatar systems, which typically consist of three stages: (i) learning priors of human appearance and motion, (ii) creating a personalized avatar, and (iii) animating the avatar. To limit the scope, we focus on the prior learning and avatar creation stages. We define current avatar representations and introduce a taxonomy that categorizes existing work along multiple axes, including body regions and employed priors. We review methods for full-body and head avatars, as well as layered representations that decompose the body into components such as hands, hair, and garments. Finally, we outline common underlying principles, reference key literature for newcomers, and discuss open challenges and future research directions.

1. Introduction

For decades, the vision of constructing a photo-realistic digital representation of a person has driven research at the intersection of Computer Graphics, Computer Vision, and Machine Learning. Richard Feynman famously stated, "What I cannot create, I do not understand," highlighting the epistemic value of creation as a means of understanding human appearance, behavior, and motion.

But why do digital humans matter? There are already several industries that depend on digital humans, especially in the entertainment field (gaming, VFX), but also in the e-commerce field (virtual

try-on). As interactions shift into virtual and hybrid environments (VR/AR, telepresence, online collaboration), people need human-centered representations that preserve identity, expression, and social nuance. *Digital humans become the bridge between physical and virtual communication, not only between humans, but also between humans and machines.*

A central question that is asked by many researchers, given the recent developments, is: *Do we need 3D avatars in a world which is increasingly dominated by 2D representations?* With the rise of image and video diffusion models, animatable portraits [XCG*24, HGZ*23], and stylized video avatars [OT24], 2D avatars have demonstrated their value. However, 3D avatars come with the promise of spatial interpretability and the ability to interact naturally within spatial environments, such as augmented and virtual reality, teleconferencing, and embodied AI. These capabilities render 3D avatars indispensable in scenarios where 2D representations fall short due to their inherent limitations in viewpoint consistency, representation compactness, render efficiency, and dynamic context adaptation.

Scope. In this report, we will review the state-of-the-art in building 3D digital humans. However, we will also discuss the nuanced relationship between 2D and 3D digital humans. Our main goal is to develop a taxonomy for the field of digital humans, which will be especially helpful for new researchers in the field. This report covers both human head avatars and full-body avatars. We also present state-of-the-art methods for individual components such as hair, hands, and garments. We emphasize topics related to building universal priors for avatars and analyze different representations used for tasks in both 2D and 3D. Note that we assume avatar motion and control signals are given, and concentrate on techniques for photorealistic appearance modeling and synthesis.

Selection scheme. This report aims to provide a comprehensive overview of the current research field, list key literature for readers exploring specific areas, and help standardize notations and definitions. The report includes papers published in the proceedings of major computer vision and computer graphics conferences from 2021 to 2025, as well as preprints available on arXiv. The papers were carefully selected to fit the scope of this survey and to provide readers with an overview of a broad range of techniques. Readers are encouraged to consult the cited publications for deeper insights.

Structure of the report. After positioning the report in relation to existing surveys in Section 2, we introduce a formal definition of avatars in Section 3. In Section 4, we introduce the fundamental components employed by modern methods to build prior models that can generalize across multiple actors, leveraging generative frameworks and statistical analysis. In Sections 5 to 9, the main chapters, the state-of-the-art methods for different body parts are presented in detail. As photo-realistic avatars also come with certain risks of misuse, we discuss social and ethical implications in Section 10. The report concludes with a discussion on the future of digital avatars in Section 11. Specifically, we formulate important research questions that are still open, such as improved efficiency, behavior modeling, spatial-aware avatars, or spatial audio.

Overall, the report follows the digital human pipeline from rep-

resentation and priors, through per-region modeling of heads, hair, hands, and garments, to full-body integration. As such, our goal is to paint a holistic picture of the current state of digital humans.

2. Related Surveys

Modeling the appearance and motion of 3D human avatars has attracted considerable research in the areas of Computer Graphics, Computer Vision, and Machine Learning. The focus of this state-of-the-art report lies on reconstructing controllable human avatars leveraging a diverse set of 2D and 3D priors. One of the most prominent priors in the field of digital humans are morphable models. Here, we want to highlight the excellent report by Egger *et al.* [EST*20]. However, general 2D priors for natural image and video generation are also widely used. We refer to the state-of-the-art reports on controllable video generation [MFH*25], as well as the report on diffusion models for visual computing [PYG*24]. We do not review literature in the field of motion capturing and motion synthesis. We want to direct the interested reader to the report on face tracking by Zollhöfer *et al.* [ZTG*18], the 3D face reconstruction survey of Morales *et al.* [MPS21], and a report on non-rigid 3D reconstruction by Tretsch *et al.* [TKB*23]. Modeling the appearance of digital humans relies on different 3D representations that we will briefly introduce in Section 4. For an in-depth review of current representations and the corresponding rendering schemes, we refer the reader to the reports on neural rendering by Tewari *et al.* [TFT*20, TTM*22]. The survey on 3D human avatar modeling of Wang *et al.* [WCHW24] is closely related to our report. However, they focus on general avatar reconstruction, pixel-aligned representations, and generative modeling, e.g., with text-based control. While the work by Gu *et al.* [GLW*25] provides a detailed survey of neural field-based techniques for reconstructing 3D human avatars, and Wang *et al.* [WCHW24] offers a thorough review of interaction design paradigms for 3D virtual humans, our survey aims to take a broader perspective. We cover a wider range of topics across the construction of digital humans, including foundational reconstruction and synthesis methods as well as specialized methods for human heads, hair, garments, hands, and full body avatars, thereby providing a more comprehensive overview of how digital humans are built and represented. In particular, our report includes a review of commonly used human priors as well as a unified framework of avatar systems facilitating classification of existing methods.

3. Definition of an Avatar System

Given the breadth of approaches covered in this report, we propose a framework for categorizing different avatar systems. We define an *avatar* as a function $\mathcal{A}(\cdot)$ that represents the appearance of a person by taking a control signal $\mathcal{C}$ as input to synthesize an image I.

Control signals. The control signal $\mathcal{C}$ consists of animation controls $\mathcal{M}$ to animate the facial expression or body pose, optional lighting $\mathcal{L}$, and camera controls π that specify the viewpoint:

$$\mathcal{C} = \{\mathcal{M}[, \mathcal{L}], \pi\}.$$

In our framework, control signals $\mathcal{C}$ are restricted to modifying non–identity-related attributes of the avatar, such as pose or rendering viewpoint. By contrast, inputs that would alter the avatar's
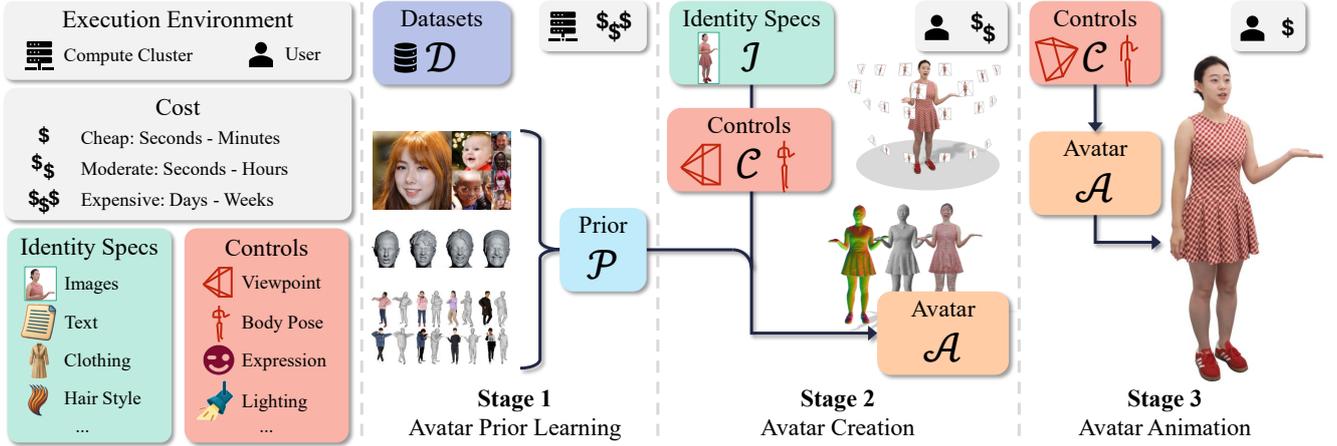
Figure 2: *A Common Framework of Avatar Systems. Images adapted from FFHQ [KAL*21], NPHM [GKG*23], THuman [YZG*21], and TaoAvatar [CHW*25].*

identity, such as editing instructions or alternative reference images, are incorporated during the avatar creation process which we will describe below. The animation controls $\mathcal{M}$ may include low-level specifications, such as per-frame body joint angles, as well as high-level cues, such as emotional states.

2D and 3D Avatars. The *avatar* can be based on an explicit or an implicit image formation process. A *2D avatar* $\mathcal{A}_{2D}$ directly synthesizes the final image I given all control signals $\mathcal{C}$:

$$I = \mathcal{A}_{2D}(\mathcal{C}).$$

Note that in our definition, a *2D avatar* is still controllable by the camera viewpoint π, which creates an apparent 3D effect without true 3D modeling. In contrast, a *3D avatar* $\mathcal{A}_{3D}$ factorizes the image formation process into two parts: (i) Creating a 3D representation $f_{3D}(\mathcal{M}[,\mathcal{L}])$ given the animation control signals $\mathcal{M}$ and optional lighting $\mathcal{L}$, and (ii) rendering the created 3D representation from camera π using a renderer $\mathcal{R}$:

$$I = \mathcal{A}_{3D}(\mathcal{C}) = \mathcal{R}(f_{3D}(\mathcal{M}[,\mathcal{L}]),\pi).$$

As such, a *3D avatar* can be seen as a special case of a *2D avatar* where the use of a 3D representation in the pipeline is enforced. An *avatar system* produces avatars matching specific identity inputs $\mathcal{I}$, such as a person's images or a textual description. Current avatar systems consist of (1) an optional avatar prior learning stage, (2) the avatar creation stage, and (3) an avatar animation stage. These stages differ by how costly they are and by whom they are executed: While learning a prior $\mathcal{P}$ on a dataset $\mathcal{D}$ may be very expensive and is only executed once by the avatar system's builder, the creation stage is executed by the end user for each desired avatar, and the animation stage is run for each final frame, ideally in real-time. A schematic overview of the different stages in an avatar system is depicted in Figure 2.

Stage 1 (optional) - avatar prior learning. The prior learning stage enables under-constrained avatar creation tasks that are unsolvable by optimization alone, for example, single-image avatar generation. A prior $\mathcal{P}$ is distilled by training a generalized model

on diverse human data, requiring a large-scale dataset $\mathcal{D}$:

$$\mathcal{D} \rightarrow \mathcal{P}. \qquad (1)$$

Publications vary widely in the types and modalities of data they use (see Table 1 - Prior Stage).

Stage 2 - avatar creation. In the creation stage, the actual avatar $\mathcal{A}$ is created from identity specifications $\mathcal{I}$, corresponding control signals $\mathcal{C}$, and (optional) human priors $\mathcal{P}$. For example, this could involve optimizing for an animatable 3D representation given videos of a person, or generating an avatar from text:

$$\{\mathcal{I},\mathcal{C},\mathcal{P}\} \rightarrow \mathcal{A}. \qquad (2)$$

The resulting avatar $\mathcal{A}$ is intended to represent the person specified by the identity parameters $\mathcal{I}$. This formulation encompasses both (i) unimodal approaches, in which an avatar is generated from images or a textual description, as well as (ii) multimodal and iterative creation processes, where editing instructions modify an existing avatar or additional specifications such as clothing materials further refine its appearance. To make avatar creation tractable, many systems use existing priors such as the geometry of a 3D morphable model. Additionally, the input requirements can vary drastically, ranging from a single image to multi-view video recordings with expensive mesh registrations (see Table 1 - Creation Stage). These requirements determine in which scenarios an avatar system can be deployed.

Stage 3 - avatar animation. The animation stage enables animation via control signals $\mathcal{C}$ and produces the final image I:

$$I = \mathcal{A}(\mathcal{C}). \qquad (3)$$

The animation stage is mainly dependent on the animation signal, but may also incorporate additional controls such as lighting (see Table 1 - Animation Stage). The animation speed is crucial in this stage, as the avatar should ideally be animated in real-time on an end-user device such as a smartphone or VR headset.

Note that while this report discusses all three stages, it does not have a focus on how the control and animation signal $\mathcal{C}$ is obtained

or generated. For face and head tracking that is used as control and animation signal in a series of methods, we refer to the report of Zollhöfer *et al.* [ZTG*18]. In addition, we give a short overview of hair and garment animation in Appendices B and C.

With this framework in place, we next introduce the fundamental building blocks, 3D representations, human priors, animation controls, and material models, that underlie the methods reviewed in the subsequent chapters.

4. Fundamentals

This section covers the fundamental concepts for creating 3D digital avatars that recur across all body regions and creation scenarios, following the progression from the avatar's representation and creation to its animation and realistic appearance. This addresses four fundamental questions:

- **What is it made of?** We begin with an overview of core 3D representations such as meshes, point clouds, and neural fields, which form the avatar's underlying structure.
- **How do we build it?** Next, we explore types of human 2D and 3D priors – statistical models crucial for generating plausible digital humans, especially from limited and incomplete data.
- **How does it move?** We then discuss the control signals and animation of digital humans.
- **How does it look?** The section concludes with an examination of commonly used material and lighting models, which determine the avatar's photorealistic appearance.

4.1. What is it made of? – 3D Representations

For the digitization of a human, we need a digital representation that stores information about the appearance of the human, including its geometry and material properties. These representations can be completely implicit and not interpretable by a human (e.g., the latent codes of some diffusion model), but they can also be human-interpretable, e.g., a mesh, and, thus, compatible with classical computer graphics concepts and pipelines. Compatibility with classical graphics pipelines enables not only explicit camera control, but for example also grants access to established geometry processing tools that allow for editing and animation. The reader is advised to consult the state-of-the-art report on 3D neural rendering [TTM*22] where a detailed introduction to the different 3D representation types is given. In the context of digital humans, we predominantly see 3D representations that either define the surface (explicitly with a mesh or implicitly with a signed distance function (SDF)) or the volume (occupancy, optical densities, radiance fields). For the surface and volume representations, different types of rendering schemes can be used. Every surface that can be indexed can be rendered in a forward process, where the surface is projected onto the image plane. Techniques like rasterization are extremely efficient, as modern GPUs have dedicated hardware units for it. Implicit representation, as well as volume information, needs to be gathered along rays that originate at a camera position and are cast through a pixel into the scene. This process is called ray casting, or ray-tracing if additional rays are shot to evaluate global illumination effects. Both ray casting and rasterization allow for explicit control of the camera.

Surface and volume representations can be stored in discrete, e.g., points, triangles, voxel grids, or continuous formats. In the past few years, continuous functions were widely used, especially in the form of neural networks, as they can be seen as universal function approximators that can store 3D content information. A very prominent example is NeRF (neural radiance fields) [MST*20], a volumetric representation that uses a multi-layer perceptron (MLP) as a function approximator to store optical density and radiance information at given points in 3D space. With a differentiable volumetric integration scheme using ray casting, this representation can be used for reconstructing and storing a digital model of a scene or object from a set of input images. The model can then be used for novel view synthesis by changing the camera pose.

Recently, many approaches have been using 3D Gaussian primitives [KKLD23]. These are radial basis functions (RBF) [CBC*01] that span a smooth function across a 3D space. It can be seen as a hybrid between an explicit representation (discrete point samples) and a smooth continuous function (Gaussian kernel) attached to it. The radial basis functions are often initialized on the surface by leveraging sparse point clouds obtained via classical structure from motion [SF16, SZPF16], learning-based dense point-cloud estimations [WLC*24a, LCR24, WCK*25], or from a reconstructed mesh (e.g., from FLAME [LBB*17]). Using a volumetric rendering scheme that leverages rasterization of the primitives, it can be implemented in a very efficient way, and, similar to NeRF, allows for novel view synthesis once the representation is optimized.

For digital humans, motions and surface deformations are a key component. In this context, explicit representations such as the Gaussian primitives have an advantage over implicit structures such as NeRF. For 3D Gaussians, deformation fields can be applied in a forward process, while implicit representations need a backward process. Such deformation fields for modeling the human motion can be based on a classical motion prior like a Blendshape-based morphable model [LBB*17, BV99], a motion proxy, e.g., cage-based [STH*24] or skeleton-based deformation (often based on linear blend skinning (LBS) [LMR*15]), physics-based simulation (especially, in the context of hair and garment), or learned representations. It is important to understand that these different representations allow for various controls of the avatar. For example, learned deformation field representations like HyperNeRF [PSH*21] or Nerfies [PSB*21] are able to model the deformations in an input sequence conditioned on a time stamp, which enables free viewpoint video playback of the sequence. However, these methods do not allow for generating novel motion. In contrast, morphable models [LBB*17, BV99] and skeleton-based deformations [LMR*15] can be controlled by the underlying motion parameters – they can be seen as generative models that could generate an infinite amount of novel deformation fields.

4.2. How do we build it? – Human Priors

While the 3D representations discussed above define what an avatar is made of, they do not by themselves resolve the ambiguities inherent in avatar creation. Many avatar creation scenarios are heavily under-constrained, such as creating an avatar from a text prompt or a few photos. In these cases, the provided identity specification is not sufficient to describe the desired avatar and therefore a prior
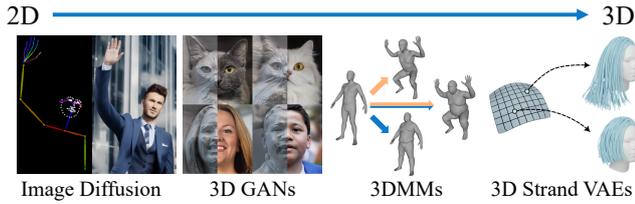
Figure 3: *Types of 2D and 3D priors commonly used for digital humans. Images adapted from Controlnet [ZRA23], EG3D [CLC*22], SMPL [LMR*15], and Perm [HSS*25].*

model encoding generic beliefs about humans has to be employed to hallucinate or fill in unseen regions. Such human prior models can be quite diverse in their design (see Figure 3). Some may only provide guidance on typical 3D human surface geometry while others focus on human appearance in 2D. All such priors are learned from large-scale datasets of human examples.

4.2.1. What is a Human Prior?

In the context of this report, a human prior models the distribution of human shape, appearance, and motion in 2D or 3D. Typical examples include 3D morphable models (3DMMs) that provide animatable geometry, e.g., by rigging 3D Gaussians to a 3DMM mesh [QKS*24]. Other approaches leverage image diffusion model priors to regularize unseen viewpoints [TDK*25, TZTL25] or employ 3D GANs for avatar reconstruction from sparse inputs [TZH*24a].

4.2.2. 3D Morphable Models (3DMMs)

Pioneered by Blanz and Vetter [BV99], there is a long-standing tradition of building mesh-based 3D parametric models for bodies and body parts [LMR*15, PCG*19, LBB*17, PKA*09, RTB17, PPM*23]. These are typically obtained by first registering a generic human template mesh to a dataset of 3D human scans and then performing dimensionality reduction techniques on the registered meshes. The result is a (linear) parametric 3D geometry model with low-dimensional controls for aspects such as body shape, skeletal articulation, facial expression, or head pose. For more details on 3DMMs, see the survey by Egger *et al.* [EST*20]. Recent works have explored different 3D representations to build 3DMMs [GKG*24, GKG*23, YTB*21, ZYHC22]. In the context of human avatars, 3DMMs are tremendously useful because they offer a coarse geometry proxy and disentangled control over human identity and articulation. A common approach is to rig 3D primitives, such as 3D Gaussians, to an existing animatable 3DMM [QKS*24, KPHT23]. As evident from Table 2, the most prominent control signals for animating a human avatar are still facial expression or body pose codes from a 3DMM.

4.2.3. 3D Generative Adversarial Networks (3D GANs)

With sufficient observations of a person, such as multi-view studio captures, avatar creation can be treated as a pure optimization problem. However, more challenging scenarios, such as single- or few-image reconstruction settings, require additional prior knowledge about the typical appearance of 3D humans to plausibly generate

regions of the avatar that have never been observed. One such prior is 3D GANs that learn a distribution of realistic 3D humans from a large 2D image dataset [CLC*22, AYS*24]. Once trained, a 3D GAN allows sampling random 3D humans from the learned distribution. The 3D prior learned by 3D GANs is particularly useful in underconstrained reconstruction scenarios, as the model's generative capabilities can fill in unseen regions and plausibly complete the final avatar. For example, several 3D portrait animation methods build on top of pre-trained 3D GANs to enable creating a 3D avatar from just a single image [TZH*24a, MZQ*23] or use 3D GANs as a synthetic multi-view data creation engine [DWR*24, YZR*24].

4.2.4. 2D Image Generation Priors

Similar to 3D GANs, pre-trained 2D image generation models help in under-constrained avatar creation tasks. However, additional measures are necessary to harness the learned 2D human appearance prior to 3D completion. Two major questions arise: (i) How to extract 3D knowledge from the 2D image generation model, and (ii) how to ensure that its output is relevant to the target avatar. Regarding (i), common approaches use the image generation model either as a critic, e.g., via Score Distillation Sampling [PJBM23, ZMF*24], or as a data generation engine that synthesizes additional images from various viewpoints, poses, or facial expressions for subsequent avatar reconstruction [TZTL25]. Another approach is to use a pre-trained image generation model as a neural renderer to directly generate the pixels of the final avatar [KGN24, DZX*23, RYV*25]. Regarding (ii), a typical solution conditions image generation on coarse 3DMM renderings or on facial landmarks of the target [TDK*25, TZTL25, ZMF*24, WPJT25]. This way, the viewpoint and other aspects are controlled in the generated images, ensuring that they provide constraints for avatar reconstruction.

4.2.5. Video Generation Priors

Similar to 2D image priors, open-source video generation models [HDZ*22, BDK*23, WWA*25, HCB*24, KTZ*24] can be a powerful tool to create avatars in under-constrained scenarios. For example, such models can be used to generate additional data of a person, which can then be distilled into an animatable 3D avatar [ZMFC25, TZT*25]. Moreover, video diffusion models with animation control [CGH*25, HGZ*23, LJY*25, MLW*24] are capable of producing highly realistic portrait animations, but typically lack precise control over camera motion. To overcome this limitation, several recent works have investigated methods for jointly controlling camera trajectories and human motion in generated videos [LXS*25, LXL*25a, CZL*25, LZL*25b]. Collectively, these advances suggest that controllable video generation models may emerge as a viable alternative to fully 3D-based avatar modeling.

4.3. How does it move? – Animation Control

Beyond representations and priors, a key aspect of digital humans is their animatability, where the animation can be driven by diverse control signals, as seen in Table 2.

3DMM expression 😈 and pose 🏃 . The predominant approaches rely on low-dimensional pose and expression parameters from

mesh-based parametric 3D morphable models (3DMMs) of the body [LMR*15, PCG*19], head [LBB*17], or hands [RTB17]. These 3DMM parameters can be integrated into the avatar's architecture in several ways. Some methods directly apply the model's Linear Blend Skinning (LBS) pose articulation and expression blendshapes to deform the avatar's core representation. This technique is used across avatar types, including full-body methods where latent codes are anchored to and deformed by a SMPL model [PZX*21], head avatars where 3D Gaussian splats are rigged to a FLAME head mesh [QKS*24], and hand [KPHT23] and clothing methods [FYP*22], where the 3DMM's articulation drives the deformation of the final geometry. More indirectly, some methods use the 3DMM for 2D-based neural rendering by rasterizing feature maps attached to the 3DMM's body [YLB*25] or head [KGN24] mesh surface from a target viewpoint, which is then input to an image generator. At a higher abstraction level, some techniques bypass the 3DMM's explicit geometry entirely and instead directly condition the avatar's neural representation on the 3DMM's expression parameters [GTZN21]. One key advantage of using a generic 3DMM is the ability to drive the avatar with off-the-shelf tools that reconstruct body and face parameters from images [FCB*21, FFBB21]. Furthermore, these models can be animated via high-level signals such as audio 🔊 involving music [TCL23] or speech [CDA*24, ASK*24] modulated by categorical emotions [DCT*23, PWS*23], or through textual descriptions [PBV23].

General expression 🎛️**.** To capture fine-grained, subject-specific details beyond the linear basis of generic 3DMMs, many methods learn custom neural embeddings. This involves creating a personalized expression latent space for heads, for instance by modeling the face as a set of deformable volumetric primitives controlled by a latent expression code [LSS*21], learning neural deformations on top of a SMPL body mesh [LHR*21], or modeling custom hand articulations on top of an LBS-articulated mesh [ISS*23] or an articulated MANO skeleton [CHV*22]. While this produces higher-fidelity results, it requires a dedicated encoder to map driving inputs (e.g., multi-view images 📷 [TKG*24], monocular videos 📹, or speech 🔊) to this custom expression space, which is trained jointly with the avatar [RBS*22] or as a separately fine-tuned regressor [ZBBT25].

Simulation 🌳**.** For avatars with secondary elements like clothing and hair, a common two-stage control mechanism uses the primary body motion (e.g., specified by SMPL pose parameters), to inform a secondary simulation. This secondary step can involve physics-based simulation using 3DMM meshes as colliders for simulating clothing layers [ZZY*24, XBS*22] or hair dynamics [ZSB*24] (see Appendices B and C).

Parametric model-free control. Finally, some methods bypass parametric models entirely, opting for direct frame-based or audio-based control. These approaches learn a direct mapping from a driving video for reenactment [dLGHT*25, TZH*24a] or from speech with control signals such as emotion or eye gaze to synthesize expressive motion without any explicit 3D model [XCG*24].

4.4. How does it look? – Material and Relighting

The final fundamental aspect concerns appearance. A key requirement for digital avatars is consistent visual accuracy under varying lighting conditions. This depends on modeling both surface material properties and scene illumination. One of the main challenges lies in properly modeling the materials of digital avatars. For example, human skin is translucent and is traditionally modeled using subsurface scattering [HK93, WJMLH01]. Garments present a similarly complex challenge due to their anisotropic structures [War92, AS00], while hair is particularly difficult to model because of its fine geometry and complex light interactions [MJC*03, ZYWK08].

This report considers two illumination configurations supported by the selected relighting methods: **distant** and **point** light sources.

Distant light 🌤️ assumes the light source to be located infinitely far away from the surface. Consequently, the incoming direction and intensity are constant over the entire object, and the illumination depends only on the incident direction ω_i (if self-shadowing and other effects are ignored). This formulation models sunlight as a directional source or environment illumination represented by a high-dynamic-range (HDR) environment map [DHT*00]:

$$L_i(\mathbf{x}, \omega_i) = L_i(\omega_i). \tag{4}$$

For Lambertian surfaces, illumination is efficiently approximated with spherical harmonics [RH01]:

$$L_i(\omega_i) \approx \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} c_{\ell m} Y_{\ell m}(\omega_i), \tag{5}$$

where the incoming direction ω_i is used to evaluate the spherical harmonic basis functions $Y_{\ell m}$.

Local light 🔦 represents one or more light sources positioned at a finite distance from the surface. Consequently, both the direction and intensity of the incoming light vary with the surface position $\mathbf{x}$. The illumination in this case is expressed as:

$$L_i(\mathbf{x}, \omega_i) = \frac{I(\mathbf{p})}{\|\mathbf{p} - \mathbf{x}\|^2} V(\mathbf{x}, \mathbf{p}), \tag{6}$$

where $\mathbf{p}$ denotes the light position, $I(\mathbf{p})$ its emitted intensity, and $V(\mathbf{x}, \mathbf{p})$ a visibility term indicating whether the light source is visible from the surface point $\mathbf{x}$. This configuration models localized illumination, such as a lamp, flash, or other nearby light sources. Further details on material representation and relighting implementations are provided in the appendix.
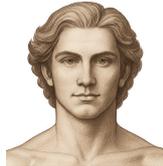
With these fundamentals established, i.e., 3D representations, human priors, animation controls, and material models, the following sections will discuss the state-of-the-art methods for individual body regions to full-body avatars, beginning with human head avatars.

Prior Stage					
Datasets	**2K** 2K2K **AI** AIST **AH** ActorsHQ **A** Ava256 **BC** BCNet **CT** CT2Hair **V** CelebV-HQ **CD** Cloth3D **CH** CustomHumans **E** Ensemble **F** FFHQ **FS** FaceScape **FV** FaceVerse **HD** Human4DiT **IH** InterHand2.6M **I** Internal **LF** LPFF **M** MEAD **MH** MRI-Hand **NH** NPHM **N** NeRSemble **NS** Next3D samples **PH** PanoHead samples **RP** RenderPeople **SR** SURREAL **SP** Sewing Pattern **TH** THuman2.0 **TW** Twindom **H** USC-HairSalon **V** VFHQ **VC** VoxCeleb2 **Z** ZJU-MoCap				
Data Type	🖼 Real 🧑‍🎨 Synthetic ⚙ Generated (e.g., by a pre-trained generative model)				
Data Modality	🦰 3D Strands 🧠 MRI ⬡ Meshes 📹 Mono video 📷 Multi-view image 📹 Multi-view video 📷 OLAT multi-view image 📹 OLAT multi-view video 🧵 Sewing Pattern 📦 Single image				
Creation Stage					
Needed Assets	**C** Cameras **HP** Hand Pose **L** Known Lighting **M** Meshes **M** Registered Meshes **T** Textures **3** Tracked 3DMM **S** segmentation masks				
Input	📄 Text 🔊 Audio △ Zero 📦 One 🖼 Few 📹 Mono video 📷 Multi-view images 📹 Multi-view video 📷 OLAT multi-view images 📹 OLAT multi-view video 🔷 Depth 🧠 3D density volume 🔵 Point cloud				
Additional Priors	**3** 3DMM **CC** ContourCraft **DF** DeepFeatures **DH** DeepMVSHair **G4** GPT-4V **GT** Garment templates **GC** GroomCap **S** Grounded-SAM **D** Image Diffusion **H** NeuralHaircut **P** Physics **PA** PortraitAnimation **R** Relighting **SS** Semantic segmentation **WV** Wave2Vec				
Creation Speed	🐌 Slow (>6 hours) ▶ Medium (<6 hours) ▶▶ Fast (<30 minutes) ▶▶▶ Instant (<1 minute)				
Animation Stage					
Animation Signal	😈 3DMM expr 🔊 Audio 🎛 General expr 📷 Multi-view image 🤸 Pose 📦 Video				
Lighting Control	🌤 Distant Light 🔦 Local Light				
Animation Speed	▶ Offline (<1 fps) ▶▶ Interactive (>5 fps) ▶▶▶ Real-time (>30 fps)				
Image Synthesis	**G** 3DGS **MV** MVP **M** Mesh **N** NeRF **NR** Neural Rendering **S** SDF **ST** Strands				

Table 1: *Avatar pipeline taxonomy legend of the three stages: prior learning, avatar creation and avatar animation.*

5. Human Head Avatars

Human head avatars play a crucial role in building a believable digital replica of a person. The human face is a key identifier of a human conveying personality, intent, and social cues. The nuanced modeling and synthesis of facial expressions and accurate likeness foster trust, empathy, and effective communication. The level of motion and appearance quality of human head avatars can differ drastically depending on the available data. The film industry is using sophisticated studios to record multi-view, multi-modal data to recover high-quality, relightable and editable avatars. However, such data is not available for end users who might want to have an avatar for a teleconference at home. Therefore, there is intense research in building avatar systems that work with sparse observations of a person, requiring powerful priors to compensate for the lack of data. We review this transition from high-quality to few-shot reconstruction methods.

5.1. Studio-level 3D Head Avatars

This category aims at creating the highest-quality virtual avatars, at the cost of requiring expensive capturing set-ups to gather enough identity observations $\mathcal{I}$, i.e., extensive calibrated and time-synchronized multi-view videos that can take hours to capture. Consequently, $\mathcal{I}$ provides dense observations of the person, w.r.t. both view angles and facial expression.

Learned expression space. While early approaches were still mesh-based [LSSS18], the field quickly evolved to volumetric representations, such as NeuralVolumes [LSS*19] and Mixtures of Volumetric Primitives (MVP) [LSS*21]. Upon the invention of 3DGS [KKLD23], 3DGS-based avatars were quickly proposed [SSS*24, TKG*24]. All these methods learn expression codes that act as motion control $\mathcal{M}$ in an end-to-end manner, yielding high flexibility, but rendering cross-reenactment difficult. GEM [ZBBT25] compresses a studio-level head avatar into a personalized low-dimensional eigenbasis. At inference time, for self- and cross-reenactment tasks, the coefficients for the eigenbasis can be regressed using a neural network, enabling real-time performance, and abandoning the need for 3DMM tracking.

3DMM-based expression space. A related line of work derives motion control $\mathcal{M}$ from 3DMM expression codes [ZBT23]. GaussianAvatars [QKS*24] attach 3DGS-primitives to the mesh faces of FLAME [LBB*17]. GaussianHeadAvatars [XCL*24] increase

Digital Human Avatars	Avatar Prior				Avatar Creation					Avatar Animation						
	Prior Dataset Size	Datasets	Data Type	Data Modality	Needed Assets	Input	Additional Priors	Req. Optimization	Creation Speed	Animation Signal	Lighting Control	Animation Speed	Image Synthesis	Image Refinement	Contents	
Face																
RGCA [SSS*24]					C M T		3	X					G			
GaussianAvatars [QKS*24]					C 3		3	X					G			
GaussianHeads [TKG*24]					C 3		3	X					G			
Neural Head Avatars [GPL*22]					C 3		3	X					M			
NeRFace [GTZN21]					C 3			X					N			
INSTA [ZBT23]					C 3		3	X					N			
URAvatar [LCS*24a]	345	I			C M T		R	X					G			
HairCUP [KSN*25]	252	I			C M T		D	X					G			
Avat3r [KRS*25]	256	A			C		DF						G			
InvertAvatar [ZSW*24]	77500	F NS V			C 3		3						N	X		
Cap4D [TZTL25]	6317	A M N V			C 3		3 D	X					G			
HeadNeRF [HPX*22]	4543	F FS I			C		3	X					N	X		
GPHM [XWZ*24]	15625	FV NH N V			3		PA	X					G	X		
GAGAvatar [CH24]	7228	V			C 3		3 DF						G	X		
VOODOOXP [TZH*24a]	103220	V F N											N	X		
VASA1 [XCG*24]	9500	I VC					PA WV						NR			
GaussianSpeech [ASK*24]					C 3		3 WV	X					G			
HeadStudio [ZMF*24]							3 D	X					G			
AnimPortrait3D [WPJT25]	89590	F LF					3 D	X					G			
Next3D [SWW*23]	70000	F					3						N	X		
Full-body																
RFGCA [WSS*25]					L 3		3	X					G NR	X		
HuGS [MSD*24]					C 3		3	X					G			
Driv. Vol. Avatars [RBS*22]					3		3	X					MV			
Driv. Aware Avatars [BWS*21]					T 3		3						M			
Neural Body [PZX*21]					3		3						N			
Anim. Gaussians [LZWL24]					3		3	X					G			
TaoAvatar [CHW*25]					C 3		3	X					G			
AvatarRex [ZZZ*23]					3		3	X					N			
D3GA [ZBS*25]					3		3						G			
GPS-Gaussian [ZZS*24]	1926	I TH						X					G			
Instant Avatar [JCSH23]					3			X					N			
Vid2Avatar-Pro [GLK*25]	1000	I			3		3	X					G			
Intrinsicavatar [WAGT24]					C 3		3	X					N			
Relightable avatar [XPG*24]					C 3		3	X					S			
RANA [ICN*23]	400	RP			C T 3		3	X					M			
Fresa [WPW*25]	1100	I			3								S			
PuzzleAvatar [XYL*24]	525	TH					3 G4 S D	X					M			
HumanRAM [YLB*25]	7100	AH HD TH Z			3								M NR	X		
ICON [XYTB22]	450	RP			3		3						S			
PERSONA [SM25]					3		3 D	X					G			
AniGS [QZZ*25]	6124	2K CH RP TH TW			3		3	X					G			
IDOL [ZLW*25]	100000	E			3		3						G			
GNaRF [BKY*22]	175	AI SR											N	X		
Hands																
RelightableHands [ISS*23]	2	I			3		3						MV			
Hand Avatar [CWS23]	27	IH			3		3	X					N			
URHand [CMG*24]	93	I			HP M T		3	X					M			
Nimble [LZQ*22]	35 /20	MH			3		3	X					M			

Table 2: *Digital Humans Taxonomy: We show representative state-of-the-art methods for human faces, bodies and hands.*
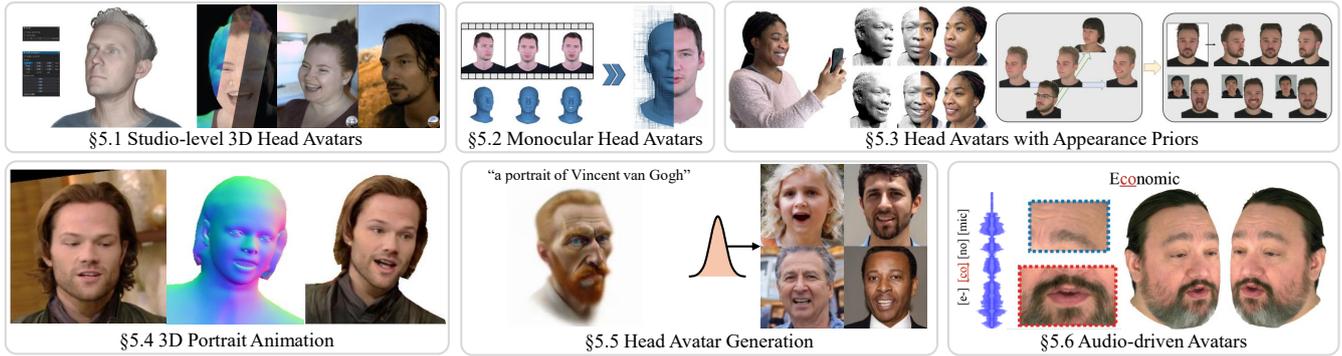
Figure 4: *Overview of different research fields in 3D Head Avatars. Representative images are taken from GaussianAvatars [QKS*24], RGCA [SSS*24], INSTA [ZBT23], Cao et al. [CSK*22], GPHM [XWZ*24], ROME [KSLZ22], HeadStudio [ZMF*24], Next3D [SWW*23], and GaussianSpeech [ASK*24].*

the dynamics by proposing MLP-based neural deformations conditioned on $\mathcal{M}$. NPGA [GKR*24] focus on higher-fidelity motion control, by replacing the mesh-based 3DMM with a neural-3DMM, i.e., MonoNPHM [GKG*24], and by introducing per-primitive features that allow for primitive-specific expression behavior.

Relightable studio avatars. Relightability is another core quality for the application of avatars in virtual environments. To this end, elaborate light-stage capture set-ups have been built that can control the lighting position, direction, and intensity for each timestep. RGCA [SSS*24] proposes a relightable formulation for 3DGS, based on pre-computed radiance transfer. The follow-up Becoming-gLit [SGN25] proposed an alternative shading formulation, which requires less diverse lighting conditions during avatar creation. Finally, DiffRelight [HCT*24] disentangles relighting from the avatar's facial animation by exploiting the image prior of a 2D image diffusion model, and fine-tuning it for relighting uniformly lit 3DGS-based avatars. The diffusion model learns complex light transport effects, such as light transmission through translucent body parts, such as ears, which are difficult to model with graphics-based shading formulations. While DiffRelight suffers from flickering between timesteps, a video-diffusion backbone could ensure temporally smooth results in the future. The follow-up work Lux-PostFacto [MHM*25] employs a video diffusion model to overcome temporal inconsistencies, and focuses on relighting arbitrary portrait videos, by first learning how to delight videos on studio light stage data, and then augmenting their relighting training data with in-the-wild videos with pseudo ground truth albedo. 3DPR [RMZ*25] addresses the complementary problem of single-image 3D portrait relighting by leveraging generative priors to synthesize plausible relighting from a single photograph.

5.2. Monocular Head Avatars without Appearance Priors

Carefully captured monocular videos can also be used to construct head avatars without relying on more priors than in the studio setting, as long as the video is easy to track and showcases diverse head rotations, which essentially imitate multi-view observations. Across recent years, several methods have been pro-

posed that explored the most promising 3D representations of their time. While very early work, like Face2Face [TZS*16] used mesh-based representations, more recent approaches have explored NeRFs [GTZN21], meshes in combination with deferred neural rendering [GPL*22], InstantNGP [ZBT23], SDFs [ZAB*22], point clouds [ZYW*23], and 3DGS [XGGZ24, SWL*24]. Compared to studio-grade avatars, such methods are mainly limited due to underlying face tracking, which is complicated due to depth ambiguity, and often struggles to accurately capture extreme expressions and head rotations. Therefore, such avatars often overfit to frontal views and simple facial expressions; thus, they typically exhibit significantly noticeable artifacts when facing stronger viewpoint changes and facial expressions. In an attempt to alleviate the need for accurate 3DMM tracking during the avatar creation stage, GAN-Avatar [KZB*24] personalizes the 3D-aware GAN EG3D [CLC*22] on person-specific data, only requiring a collection of images and camera parameters. Animation is enabled by learning a mapping from 3DMM codes to the GAN latent space.

5.3. Monocular Head Avatars with Appearance Priors

While the methods above rely on carefully captured monocular videos, many practical scenarios provide even less data. As observations $\mathcal{I}$ become sparse, such as casually captured or in-the-wild monocular videos, or when provided with a few casual selfie images, the need for a strong prior $\mathcal{P}$ increases.

3D prior + fine-tuning. A popular approach for a learning-based 3D prior is Universal Prior Models (UPMs) [CSK*22, LCS*24a, MDT*25], which are UNets, operating in UV-space to transfer coarse UV-space geometry and texture from a tracked mesh to high-fidelity volumetric representation, such as MVP [CSK*22] or Gaussian primitives [LCS*24a, ZGL*25]. While UPMs produce high-quality results once a tracked mesh has been obtained, their avatar creation stage commonly includes fine-tuning of the encoder and decoder networks based on $\mathcal{I}$, which can take up to several hours [CSK*22, LCS*24a] or several minutes [ZGL*25]. In general, since UPMs operate in UV-space, they are reliant on high-quality mesh-based tracking, which is a challenging task on its own.

Photo-realistic 3D head models. Pioneered by Head-NeRF [HPX*22], some works aim at building a parametric 3D head model with an additional photorealistic appearance space. To achieve this, one has to go beyond the geometrical perspective of classical 3DMMs [PKA*09, LBB*17] and train a morphable model on a large dataset of human videos. Instead of meshes or SDFs, as common for 3DMMs, photo-realistic 3D head models therefore rely on NeRF [HPX*22, YZM*24, ZZSC22, YBM*24] or 3DGS [ZWL*25, XWZ*24, MDT*25]. The most common architectural choice is an autodecoder [PFS*19] that optimizes separate codebooks: one for each person, facial expression, or light configuration. The disentanglement of these factors often follows from the use of an (OLAT) multi-view image dataset where the same person, expression, and potentially lighting configuration is seen from multiple cameras. Once trained, the parametric head model can be fitted to any number of observations, similar to classical 3DMMs. As such, photo-realistic 3D head models are very versatile and can be used in many scenarios, from creating an avatar with a single image to multi-view video fitting. On the downside, the creation involves potentially costly fitting and assumes that the target identity actually exists in the model's latent space. An assumption that barely holds given the small identity coverage of existing multi-view datasets. While HeadNeRF [HPX*22] finds that including monocular images from the FFHQ dataset [KLA19] in their Autodecoder training leads to better generalizability, One2Avatar [YBM*24] observes that such an approach leads to limited 3D completeness in the final avatars. In general, how to train parametric 3D models on partial data, such as monocular videos or single images, remains an open problem.

Feed-forward 3D prior. Following the recent success of multi-view transformers for 3D reconstruction [WLC*24b, HZG*24, WCK*25], Avat3r [KRS*25] proposes a formulation that lifts four input images into a set of canonical Gaussian primitives using a transformer-based architecture similar to GS-LRM [ZBT*24]. Expression control is achieved using a cross-attention module that attends to latent expression codes. As such, Avat3r targets avatar creation with minimum requirements on observed images $\mathcal{I}$ and minimal compute budget during the avatar creation stage, which merely involves a single forward pass of the transformer. It is worth noting that the output is not a structured representation (e.g., a mesh or primitives attached to it), but pixel-aligned 3D Gaussian primitives.

Image-based priors. Inspired by the recent success of image diffusion and multi-view diffusion models, CAP4D [TZTL25] and GAF [TDK*25] propose a 3DMM-conditioned multi-view diffusion model, which learns to generate high-quality 3D consistent images, based on the prompted camera view and 3DMM expression. In this way, the diffusion model is tasked with generating novel views from unseen camera angles and unseen facial expressions, which are included in the optimization of a 3DGS-based avatar.

5.4. 3D Portrait Animation

Portrait animation methods further reduce the required identity specification $\mathcal{I}$ to a single image. As such, the problem becomes heavily under-constrained, and consequently, portrait animation methods are heavily prior-reliant and learning-based.

Feed-forward prediction. A straightforward approach to portrait animation is to train a feed-forward network to predict 3D Gaussian primitives, similar to Avat3r [KRS*25], but based on a *single* input image instead. Notable methods like GAGavatar [CH24] and LAM [HGY*25] predict Gaussian parameters, which are attached to a 3DMM mesh, allowing for animation of these Gaussians, but limiting the representational capacity w.r.t. expressions to that of the underlying 3DMM. While LAM purely relies on mesh-attached Gaussians, GAGavatar uses a combination of mesh-attached Gaussians and a set of pixel-aligned Gaussians, which are non-animatable. VoluMe [dLGHT*25], in contrast, purely relies on pixel-aligned Gaussians, but predicts two Gaussians per pixel, to accommodate for occluded regions. VoluMe cannot be animated via some abstract expression conditioning, but only using images from the same person, since each input image is directly transferred to 3D. While this leaves the method without a proper avatar animation stage, it demonstrates that their avatar creation stage is efficient enough for real-time performance and is temporally consistent. Earlier works [KMC24, YFZ*23, YZR*24, TZH*24b, TZH*24a] adopted the efficient TriPlane representation [CLC*22] for fast volumetric rendering, which is crucial when training on large video datasets. A technique is to predict two TriPlanes from the input image: one that contains the canonical 3D face of the person, and a second that contains the 3D motion of the driver [YZR*24].

3D GANs. Yet another approach to portrait animation is 3D GANs, with added control for facial expressions. 3D GANS [CLC*22, KGT*24, BMH*25] have the unique ability to learn a distribution over 3D heads from large 2D image collections, without the need for costly 3D datasets, which usually cover a much narrower data distribution. However, such generative models, by default, cannot disentangle identity and expression variation. Popular approaches to add expression control into 3D GANs are generating features in UV space to inherit the underlying animation capabilities from a 3DMM, as explored in Next3D [SWW*23], or training a neural network to canonicalize the generated triplanes and learning to add expressions to the triplanes using a separate network, as proposed in VOODOOXP [TZH*24a]. A single input image can be inverted using pivotal tuning [RMBCO22], resulting in a latent that can be decoded into a 3D head avatar. Another approach to add disentangled expression control to an existing 3D GAN is via latent space exploration. Here, the assumption is that the 3D GAN's latent space contains vectors representing the same identity with different facial expressions. Disentangled animation can therefore be achieved by training a separate network that maps identity and expression codes into the latent space of the 3D GAN [MZQ*23].

Image-based prior + optimization. There exist several approaches that exploit the image prior of existing image/video diffusion models to hallucinate novel views, conditioned on a single image. The synthesized novel views can be baked into a 3DGS representation to obtain a 3D avatar. CAP4D [TZTL25] and GAF [TDK*25] also support the single-image setting, while Zero-1-to-A [ZMFC25] specializes in portrait animation, by fine-tuning a video diffusion model to generate novel trajectories, with potentially higher 3D consistency.

2D and 2.5D avatars. A core question in the community is whether the avatars necessarily have to follow some underlying 3D representation, whether they can be purely 2D, or some mixture which has some 3D components. While recent trends in video and world generation suggest that a surprising amount of consistency can be achieved by scaling up 2D generative models, the specific requirements for head avatars in teleconferencing systems, such as real-time performance, long-term consistency, support for multiple avatars, and consumer-grade edge devices, add more complexity to the debate. Face-Vid2Vid [WML21] and Live-Portait [GZL*24], for example, choose to learn in a 2.5D space, using a simple orthographic projection and learned implicit 3D keypoints, which condition a warping in a 3D feature volume. Mega-Portraits [DCK*22] alleviates the bottleneck of using sparse keypoints to condition the warping by replacing them with learned latent codes. Another line of work attempts to exploit the excellent image generation properties of diffusion models. A common choice to condition the diffusion process on the desired facial expression and camera view are renderings from a 3DMM, for example, see [DZX*23, PZS*25, TZTL25, TDK*25]. As such, these approaches abandon all sorts of internal 3D representation or rendering mechanisms. Instead of relying on human-crafted 3D inductive biases, they embrace a more data-driven approach. Alternatively, some works decide to condition on landmarks [MLW*24, WWJ*25], or learnable 3DMM-attached features [GZL*25], rather than rendered 3DMM meshes. An even more extreme version is video-driven portrait animation models, which condition video diffusion models on a source image $\mathcal{I}$, and use a driving video as motion control $\mathcal{M}$ [CGH*25, CLZ*25, RYV*25, SZG*25]. As a consequence, camera and motion control in the animation stage cannot be directly modified.

5.5. Head Avatar Generation

Text to 3D head avatars. By further reducing the information content in $\mathcal{I}$, we arrive at the task of creating 3D head avatars conditioned merely on text. While native 3D generative models text-conditioned head avatars are still rare, especially when requiring animation controllability, several optimization-based approaches exist [ATDN23, ZFK*24]. For example, HeadStudio [ZMF*24] uses score distillation sampling (SDS) to optimize for Gaussian primitives rigged to the FLAME mesh. Similarly, AnimPortrait3D [WPJT25] use Interval Score Matching (ISM) and SDEdit (SDE) based on a 3DMM conditioned 2D diffusion model, inspired by Joker [PZS*25].

An alternative for text-conditioned head avatar creation is to employ a text-conditioned 2D generation model, such as FLUX [LBB*25], to create an image based on a textual description and subsequently feed the image to one of the above-mentioned portrait animation methods, in order to create an avatar.

Unconditional generation. Finally, some works aim at learning the entire distribution of 3D head avatars, enabling unconditional generation without any identity specification. Compared to the previous sections, this area is slightly underexplored. Controllable 3D GANs, such as Next3D [SWW*23], GNARF [BKY*22], or GAIA [YLS*25], are one of the few approaches for unconditional avatar generation. To achieve this, a common approach is to start with an existing static 3D GAN such as EG3D [CLC*22] and incorporate the motion prior from a 3DMM such as FLAME [LBB*17] to make it animatable. RodinHD [ZCW*24a] is a latent diffusion model to generate 3D heads, but does not support animation capabilities. However, similar to text-conditioned generation, one can employ an unconditional image generator, followed by a portrait animation method.

5.6. Audio-Driven Avatars

Many applications of head avatars are centered around speech. Thus, there is an increased interest in the physical plausibility of mouth movement during talking and proper lip synchronization. GaussianSpeech [ASK*24] builds studio-grade 3DGS-based avatars, which are animated using audio signals. VASA-1 [XCG*24], on the other hand, is a latent diffusion model that operates in the latent space of Megaportraits [DCK*22]. The identity component of the avatar is obtained using the Megaportraits encoder, and the generation task boils down to the generation of a sequence of expression codes, conditioned on an audio sequence as motion control $\mathcal{M}$. UniGAHA [TRM*25], similarly, employs a latent diffusion model to map an audio sequence to the latent expression space of a UPM.

6. Hair

Most 3D head avatar methods mentioned in the previous section treat the head as a single, monolithic representation. However, the head is composed of structurally distinct elements, most notably hair, whose geometry, dynamics, and appearance differ fundamentally from those of the face. While the head can be approximated as a smooth, largely rigid geometry, hair consists of thousands of thin, flexible fibers with complex appearance and dynamic behavior. A unified representation conflates these characteristics, limiting generalization, reducing rendering fidelity, and restricting editability. Disentangled representations, on the other hand, model hair as a separate layer from the head [KSN*25, LDN*25, OBT25, HLK*25]. Although this approach introduces additional complexity compared to embedding hair directly on the head, it offers key advantages, including accurate geometric modeling, independent control of hair and head, smooth interpolation, and compositional flexibility for layered avatar construction.

In the earliest days of computer graphics, hairstyles were created almost entirely by hand, using textured polygonal strips (hair cards) or a few manually drawn spline curves duplicated to suggest a full head of hair [may, Com18]. However, automatically reconstructing realistic hairstyles from observations is also highly challenging. A typical hairstyle consists of 100,000–200,000 individual strands, yet only a small fraction are directly visible due to severe self-occlusion. Since a strand is essentially a 3D curve, recovering the hidden internal structure is particularly difficult. Even with multi-view data, ambiguities persist, as different internal arrangements can produce nearly identical external appearances. Moreover, hairstyles must capture both global coherence, such as con-

Digital Human Assets	Assets Prior				Assets Creation							
	Prior Dataset Size	Datasets	Data Type	Data Modality	Needed Assets	Input	Additional Priors	Creation Speed	Representation	Simulation Ready	Lighting Control	Contents
Hair												
CT2Hair [SSW*23]					3			🐌	ST	X		
NeuralStrands [RSW*22]		I			3			🐌	ST	X		
GroomCap [ZCW*24b]								🐌	M ST	X		
Neural Haircut [SCD*23]	343	H			3			🐌	ST			
Monohair [WYK*24]	2 744	H					DH	▶	ST			
GroomLight [ZCV*25]							GC	🐌	G ST	X	☀️	
DiffLocks [RWF*25]	40 000	I						▶▶	ST	X		
HAAR [SZH*23]	9825	CT I H			3		H	▶▶▶	ST	X		
3DGH [HLK*25]	70000	PH							G			
Doubly [CCH*25]	658	H						▶▶▶	ST	X		
Garment												
Neural Tailor [KL22]	22 000	SP						▶▶▶	M	X		
Rec [QCZ*23]					3		GT SS	🐌	S			
BCNet [JZH*20]	48 000	BC			3			▶▶▶	M			
SCARF [FYP*22]					3 S		3	🐌	M N			
Gaussian Garments [RGW*25]					M		3 CC P	🐌	G M			
SimAvatar [LYDM*25]	20 000	CD SP					3	▶	G			

Table 3: *Digital Human Assets Taxonomy: We show representative state-of-the-art methods for human avatar assets like hair and garments.*

sistent flow directions, and local irregularities like curls, wisps, and frizz to achieve realism and physical plausibility.

In the following sections, we provide a detailed discussion of hair reconstruction and generation methods. For hair animation and simulation, as well as physics-based rendering and relighting of hair, we refer the reader to Appendix B.

6.1. Hair Reconstruction

We review the state-of-the-art hair reconstruction methods by input modality, progressing from costly multi-view systems toward affordable, consumer-level phone captures. A common strategy in image-based approaches is to optimize or regress a 3D hairstyle such that its projection matches the observed 2D orientation features in the image (often using Gabor filters). Regardless of the input modality, most hair reconstruction approaches can be grouped into two main categories: (i) *hair growing* methods and (ii) direct *hair map* prediction / optimization. Both types typically begin by estimating a 3D volumetric representation of the hairstyle along with an orientation field. *Hair growing* methods [LLR13, CWW*13, ZCW*17, ZWW*18, SHM*18, ZZ19, NWKS19, WYY*22, KCF*22, ZJL*23, WYK*24] employ a hair growing procedure by sampling root positions on the scalp and tracing strands along the orientation field. In contrast, methods such as [ZHX*18, RSW*22, SCD*23, ZCP*23, SZH*23, ZSB*24, LOZ*24, TTS*24, RWF*25] directly predict or optimize a (latent) *hair map* based on orientation fields, volumetric or silhouette information, to recover the final strand-level geometry.

6.1.1. Hair Reconstruction from Multi-view Data

Even with multi-view calibrated data, state-of-the-art methods rely on strong priors to capture the internal structure of hair. Zhang *et al.* [ZCW*17] propose a four-view image-based method, reconstructing realistic 3D hair by combining database retrieval with texture-derived 3D direction fields, even when the input views come from unpaired hairstyles. DeepMVSHair [KCF*22] introduces a deep implicit formulation, representing hair as a continuous 3D growth direction field. It predicts occupancy and direction at any query point from multi-view features, using a view-aware transformer to aggregate information and preserve high-frequency details. A hair-growing and deformation stage further refines strand geometry. Building on the hairstyle prior of DeepMVSHair for capturing accurate internal geometry, MonoHair [WYK*24] reconstructs external hair geometry from monocular video using Patch-based Multi-View Optimization (PMVO), then applies a hair growth algorithm to resolve directional ambiguities.

Nam *et al.* [NWKS19] combine a multi-view stereo method with a novel cost function, clustering, and strand growing. While their method achieves accurate reconstructions, the strands are often disconnected from the scalp and hairstyles may contain holes, limiting applicability in downstream simulation. Neural Strands [RSW*22] extends this line of work to full hairstyle reconstruction by optimizing geometry in a latent hair-map space aligned with a scalp UV map. At each iteration, strand roots are sampled and decoded using a pretrained strand prior, then projected back into 2D space via soft rasterization of lines. Geometry is supervised using RGB images, Gabor features, and 3D strand priors from Nam *et al.* [NWKS19].

Since calibrated multi-view capture is impractical in consumer applications, recent research has shifted toward monocular 360° video under natural conditions. Neural Haircut [SCD*23] introduces a two-stage pipeline for strand-based hair reconstruction from monocular video captured with a smartphone. It directly optimizes for a latent hair map using a diffusion-based hairstyle prior, which is trained on synthetic data. To establish data terms of the optimization, Neural Haircut uses a Structure-from-Motion (SfM)-based 3D geometry estimation of the visible surface and surface orientation fields extracted from the input images. To improve internal strand flow, Dr.Hair [TTS*24] introduces a Laplace-based initialization, connectivity-aware reparameterization, and an enhanced renderer. Gaussian Haircut [ZSB*24], GaussianHair [LOZ*24], and HairGS [PNK25] propose to represent strands as sequences of cylindrical 3D Gaussian primitives anchored at segment midpoints, which enables efficient rasterization and improves reconstruction quality. To address motion blur during phone capture and inaccuracies in camera poses estimated by COLMAP [SF16, SZPF16], Gaussian Haircut [ZSB*24] further proposes camera pose finetuning, leading to improved reconstruction quality. In contrast to these prior-driven approaches at the strand or hairstyle level, GroomCap [ZCW*24b] learns a neural implicit representation for 3D occupancy and orientations using a volumetric 3D orientation rendering algorithm, coupled with 2D orientation distribution supervision. Similar to [ZSB*24,LOZ*24], they use a 3D Gaussian primitives-based representation enabling direct photometric supervision from input images. It is worth noting that these methods can only handle static geometries. There are only a few methods like Hu *et al.* [HBLB17] that tackle both reconstruction and the estimation of physical hair properties, thereby bridging hairstyle modeling with dynamic simulation.

6.1.2. Hair Reconstruction from a Single Image

Reconstructing strand-based hairstyles from a single image is inherently ill-posed, as only the frontal view is observed while multiple plausible back views may exist. Furthermore, Gabor features capture local hair orientations but remain undirected, leaving directional ambiguities unresolved. To address these challenges, early methods [CWW*12, CWW*13, HMLL14, HMLL15] incorporate user-provided strokes as additional constraints. Methods like [CSW*16, HSW*17, ZWW*18] improve automation by retrieving hairstyles from a dataset, which are then refined through optimization. Recent approaches, such as Hairmony [MCH*24] employ learned hairstyle classifiers to guide reconstruction. Building on these approaches, recent single-view methods leverage learned hairstyle priors. These priors are trained on synthetic datasets of realistic strand-based hair. To mitigate the domain gap (sim-to-real), methods use orientation maps as input [ZHX*18, SHM*18, ZZ19, WYY*22, ZJL*23]. Hairstep [ZJL*23] additionally proposes dense direction maps labeled on a small dataset. In DiffLocks [RWF*25], procedurally generated curly hairstyles are used to train a diffusion model, conditioned on DINOv2 [ODM*23] features extracted from input images, to synthesize realistic strand-based reconstructions. In contrast to these complex prior models, there are also classical PCA-based hair model priors as used in Perm [HSS*25], representing both coarse and detailed hair geometry. It reconstructs hairstyles from single images via inversion in
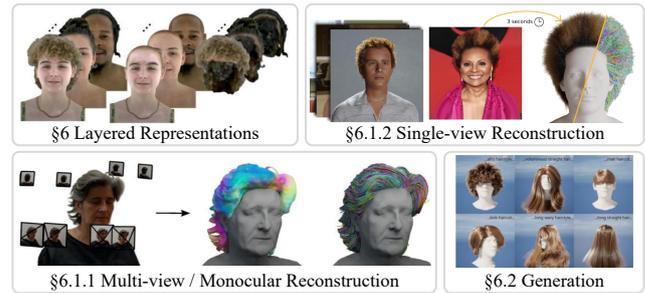

Figure 5: *Hair reconstruction methods. Representative images are taken from HairCUP [KSN*25], Difflocks [RWF*25], Neural Haircut [SCD*23], HAAR [SZH*23].*

PCA space. Building on this idea, Im2Haircut [SZP*25] trains a transformer-based prior to predict the PCA map from an input image, followed by inversion in the prior space, improving efficiency and accuracy.

The methods above aim to develop general solutions for a wide range of hairstyles; however, others focus on specific styles, such as braids or curls, using specialized priors or procedural rules. This reflects a trade-off between versatility and precision: general methods handle diverse hair but may struggle with complex structures, whereas specialized approaches achieve high fidelity for targeted hairstyles. For example, Hu *et al.* [HML*14] reconstruct braided hairstyles from a single RGB-D camera using compact procedural braid models. Luo *et al.* [LLR13] reconstruct coherent hair wisps from still images by detecting local structures, reasoning about connectivity and direction, and synthesizing strands robust to occlusions, suitable for animation. Chai *et al.* [CLS*15] further reconstruct high-quality 3D hair from minimal input by combining depth cues, a 3D helical hair prior, and Shape-from-Shading normals.

6.1.3. Hair Reconstruction from CT Data

Unlike the image-based methods that we discussed above, which can only capture the visible outer surface, computed tomography (CT) scans enable reconstruction of internal hair structures. CT2Hair [SSW*23] is the first method to leverage CT scans to generate high-fidelity 3D hair models from real-world hair wigs. The approach employs a coarse-to-fine process: guide strands with 3D orientation fields are first recovered, then dense strands are populated via neural interpolation and refined to match the density volumes. While this is an interesting experiment, CT scans for avatar reconstruction are not feasible in practice.

6.2. Hair Generation

Most hair capture techniques focus on recreating a single subject, whereas generative hair models aim to represent and manipulate a wide range of hairstyles, but have so far been less studied.

6.2.1. Unconditional 3D Hair Generation

GroomGen [ZCP*23] is the first generative model for dense, strand-level hair geometry, employing hierarchical latent spaces to represent individual strands, sparse guide hairs, and full hairstyles.

Trained on an augmented, artist-designed strand-based dataset, the model can generate realistic and diverse hairstyles at inference. Perm [HSS*25] introduces a PCA-based representation for hairstyles. Unlike prior work that jointly models global structure and local curls, Perm disentangles them in the frequency domain, separating hair geometry into low-frequency guide textures and high-frequency residual textures. These components are parameterized with separate generative models, allowing Perm not only to reconstruct hairstyles accurately but also to generate novel hairstyles, perform hairstyle editing, and serve as a prior for downstream tasks. Another hierarchical strand-level approach is proposed by Chen *et al.* [CCH*25], which first generates coarse guide hairs and then dense strands with high-frequency details. By leveraging a discrete cosine transform to separate global structure from local curls and k-medoids clustering to select guide strands, this method preserves hairstyle characteristics and supports off-grid modeling for more realistic dense strand generation.

Extending from hairstyle-level to full-head generation, 3DGH [HLK*25] models 3D human heads with composable hair and face components. Hair is represented using deformable 3D Gaussians within a template-based Gaussian splatting framework, capturing geometric variations across hairstyles. A dual-generator 3D GAN with cross-attention models hair-face correlations, enabling high-quality 3D hairstyle synthesis and composable hairstyle editing, outperforming prior 3D GAN approaches.

6.2.2. Text-based 3D Hair Generation

HAAR [SZH*23] proposes the first strand-based generative model that produces 3D hairstyles from textual descriptions. It leverages VQA-annotated [LLLL24, LLWL23] synthetic hair models to train a latent diffusion model in a common hairstyle UV space, enabling realistic and diverse hairstyle generation. Later, Strand-Head [SCT*25] extends this approach to model disentangled full head avatars. Instead of relying on large-scale paired data, they propose to distill knowledge from a pretrained 2D diffusion model to generate an avatar.

6.2.3. 2D Hair Appearance Generation with 3D Control

ControlHair [LLZ25] combines 3D hair reconstruction and physics-based animation with a conditional video diffusion model. Specifically, an appearance prior in terms of a video diffusion model is learned on real videos by fine-tuning a Control-Net [ZRA23]-like architecture based on WAN2.1 [WWA*25]. As a control signal, it leverages a reference image of the subject and orientation maps that can be extracted from original videos during train-time [ZJL*23]. At test-time these orientation maps are provided by a classical computer graphics rendering scheme that leverages 3D hair strands that are recovered from the reference image by DiffLocks [RWF*25] and animated with Blender. While the usage of such a learned hair rendering prior results in promising results, it heavily relies on the accuracy of the 3D hair strand reconstruction and the simulation. Automatically recovering physics parameters of a hairstyle from a reference image that can be used for simulation is an open challenge, and thus, ControlHair requires manual fine-tuning of the physics parameters in Blender.

PhysAnimator [XZJJ25] is a related method that follows mainly the same concept, but uses a mesh-based representation to model dynamics and contour-lines as guiding signal of the video diffusion model. It is more general than ControlHair, but lacks strand-level control and is only evaluated on cartoon animation.

6.3. Facial Hair Reconstruction

Recently, several works have focused on modeling beards, eyebrows, and eyelashes using strand-based representations. Beeler *et al.* [BBN*12] present a method that reconstructs various facial-hair styles along with the underlying skin surface. Using multi-view images, it detects and traces hairs for beards and eyebrows and lifts them into 3D to produce strand-based geometry, while also handling hair occlusion to generate realistic skin geometry underneath. Li *et al.* [LL24] further propose dense facial hair reconstruction and tracking. By combining line-based multi-view stereo and line segment matching, it recovers a dense hair point cloud, extracts strands via a forward Euler method, and tracks them over time with space-time optimization. The refined skin mesh produces highly accurate and complete facial hair models suitable for dynamic facial performances. EMS [LJZ*23] reconstructs 3D eyebrows from a single image by modeling them as fiber curves. It uses three modules: RootFinder for root locations, OriPredictor for 3D growth directions, and FiberEnder to stop growth. Trained on a synthetic 3D eyebrow dataset, EMS handles a wide range of eyebrow styles and lengths. Kerbiriou *et al.* [KAM24] introduce the first data-driven semantic model of eyelashes. Using multi-view 3D reconstruction, they build the first 3D eyelash dataset and extract semantic features to describe shape and variation, enabling intuitive and realistic eyelash synthesis for avatars and synthetic data generation.

7. Hands

Although hands are part of the body, we pay special attention to hand modeling, as hands play a central role in everyday interactions with hands, other body parts, and objects. In addition, hand modeling is uniquely challenging because it needs to support large articulations. Changes in poses lead to drastic geometry and appearance change due to large self-contact and shadow casting. In this section, we review the literature on hand priors and hand modeling, and discuss hand avatar control.

7.1. Hand Priors

Due to the large articulation of hands, various priors are proposed for hand modeling and tracking. Earlier work utilizes a collection of primitives to represent hands [RK94, OKA11a, TST*15]. Sridhar *et al.* instead use a collection of 3D Gaussians [SOT13, SMOT15]. Sphere-meshes [OKA11b, TPT16] are also used to simplify the shape representation of hands for tracking. Schmidt *et al.* [SNF14] combine voxelized primitives and signed distance function (SDF) to further improve collision handling. While these representations are useful to handle heavy collision and occlusion in an efficient manner, their geometric fidelity remains limited.

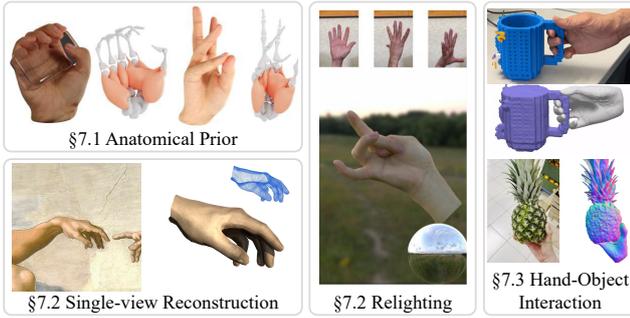A triangular mesh has gained the most popularity for shape prior

§7.1 Anatomical Prior

§7.2 Single-view Reconstruction

§7.2 Relighting

§7.3 Hand-Object Interaction

Figure 6: *Hand reconstruction methods. Representative images are taken from NIMBLE [LZQ*22], Handy [PPM*23], UR-Hand [CMG*24], and HOLD [FPK*24], respectively.*

due to its efficiency and fidelity. Similar to face 3DMMs [BV99], Khamis *et al.* [KTS*15] apply PCA to the scans of 50 people using a depth sensor to obtain a linear identity shape space. However, pose-dependent deformation is limited to linear blend skinning (LBS), leading to limited fidelity. Later, in the same spirit as SMPL [LMR*15], MANO [RTB17] learns identity linear shape space and pose-dependent corrective space independently using PCA from approximately 1000 scans consisting of up to 51 poses from 31 subjects. Non-linear variants of identity and pose correctives have also been explored. Dkulon *et al.* [KWG*19] use graph convolution to represent correctives. However, this work models identity shape and pose-dependent deformations into a single latent space, not supporting disentangled control. DeepHandMesh [MSL20] present a self-supervised method to learn non-linear identity and pose correctives parameterized by MLPs. The model is learned from 3D scans obtained from InterHand2.6M [MYW*20], showing superior details compared to the MANO model. UHM [MXJ*24] additionally incorporates image-based supervision through appearance modeling. Incorporating anatomical priors has also been explored in several works [LZQ*22, WMB19, ZWHB22]. These approaches use magnetic resonance imaging (MRI) to obtain underlying tissues and bones. These anatomical priors and physics-based volumetric prior [SWW*20] are highly effective at handling large deformations and self-contacts of hands.

Since hands share a common topology across identities, building an appearance prior on UV space is a common choice. HTML [QWM*20] builds a PCA linear appearance space of textures on UV space under a uniformly lit environment from 51 subjects. NIMBLE [LZQ*22] learns PCA linear appearance space parameterized by PBR reflectance maps (i.e., surface normal, albedo, and specular map) from 38 scanned subjects. Despite being compact, a linear basis tends to blur the final outputs. Handy [PPM*23] learns a generative model of hand geometry and appearance using StyleGAN3 [KAL*21] from approximately 1200 high-quality dome-captured data. The expressiveness of StyleGAN enables more photorealistic texture synthesis. URHand [CMG*24] builds a relightable appearance prior on UV space from 93 identities with approximately 42,000 frames of time-multiplexed illuminations for each identity. URHand does not contain a latent bottleneck unlike the aforementioned approaches, and instead directly

translates a mean texture map into relightable features. This also supports more detailed appearance. Unlike UV space, directly representing geometry and appearance in 3D space using Neural Fields [XTS*22] such as NeRF is also a promising direction due to its expressiveness and photorealism. LISA [CHV*22] learns a prior model of part-based SDF from InterHand2.6M in an end-to-end manner. OHTA [ZWS*24] presents a NeRF-based prior model that learns albedo, shadow, and opacity fields from InterHand2.6M [MYW*20] dataset. The fields are conditioned by a global latent code, and can be quickly personalized via inversion. GaussianHand [HLL*24] adopts 3D Gaussian Splatting [KKLD23] for geometry and appearance representations, and builds a prior in UV-aligned feature space from InterHand2.6M.

7.2. Hand Avatar Creation

Hand avatar creation has been explored with various input setups. On one hand, multi-view dome capture data is used to push the envelope of fidelity in hand avatars. These approaches are typically person-specific and have minimal use of priors and preprocessing, with a training time longer than a day. RelightableHands [ISS*23] builds a person-specific relightable hand model learned from multi-view dome data with time-multiplexed illumination patterns. RelightableHands directly learns a mapping from given lights to final appearance, achieving photorealistic relighting with complex global light transport effects such as subsurface scattering and non-local shadows. LiveHand [MWH*23] and HandNeRF [GZW*23] learn a person-specific non-relightable hand model represented by NeRF from multi-view dome data. LiveHand parameterizes NeRF using UV coordinates on MANO and height fields, following Neural Actor [LHR*21], whereas HandNeRF parameterizes it in a canonical space and deforms it via LBS-guided deformation fields.

On the other hand, if the input is a monocular video, the problem often becomes ill-posed due to limited observations. To address this, leveraging prior models discussed in Section 7.1 is critical. HARP [KPHT23] uses meshes as a shape representation and introduces tailored loss functions with various shape regularization terms to stabilize the optimization process. HandAvatar [CWS23] learns a photorealistic hand avatar represented by part-wise occupancy fields together with shading and albedo appearance fields. The key to the successful optimization is to condition the neural fields with a high-resolution version of the MANO model, MANO-HD. UHM [MXJ*24] and URHand [CMG*24] both constrain the geometry with a global latent code as an information bottleneck, allowing us to quickly optimize via inversion.

Similarly, having a compact latent code for geometry and appearance enables hand avatar creation from a single image. Given a single image, we can run inversion to obtain a corresponding latent code by minimizing image-based losses [QWM*20, CHV*22, ZWS*24, HLL*24]. Handy [PPM*23] uses similar loss functions to train a regressor to enable feed-forward reconstruction of a personalized hand avatar.

7.3. Hand-Object Interaction

Recent advancements have shifted toward the holistic modeling

of hands in context, specifically their interactions with hands, objects and other body parts [SGPT23, FPK*24, CMT*25, OLS*25]. HOLD [FPK*24] addresses the joint reconstruction of a hand avatar and an object from monocular video, demonstrating that simultaneous optimization reduces depth and scale ambiguities inherent in isolated modeling. Beyond object manipulation, Decaf [SGPT23] explores complex face-hand interactions by explicitly learning contact regions and local deformations from monocular images. While Decaf relies on an iterative optimization-in-the-loop to resolve final geometries, InteractAvatar [CMT*25] advances the state-of-the-art by moving toward a feed-forward approach. It achieves high-fidelity results by jointly regressing 3D Gaussian correctives that capture the fine-grained geometric and photometric nuances induced by face-hand contact.

7.4. Hand Avatar Control

Unlike hair or clothing, hand shape and appearance are mostly deterministic based on input poses. Therefore, the aforementioned hand avatar models are all solely driven by hand pose parameters. For relightable hand models [KPHT23, LZQ*22, ISS*23, CMG*24], we also take a target illumination as an input condition.

7.5. 2D Generative Models for Hands

Early diffusion-based animation models typically suffer from precisely reconstructing hand shapes [MFH*25, Hu24]. Several works focus on specifically addressing the hand quality issues of image/video diffusion models by introducing region-aware losses [FYA*24] or 3D hand embeddings [NBC*24]. FoundHand [CMZ*25] achieves precise pose and camera control of hand images with an image-to-image diffusion model.

8. Garments

Garments play an important role in building immersive human avatars. Subtle movements in body pose can have a large influence on the motion and appearance of the garment, e.g., wrinkles get visible or disappear during animation. In contrast to faces, or minimally clothed human bodies, there are no general morphable models that could be used to represent a large variety of clothing types and their deformations. Challenges like topological changes need to be handled, as well as 'free-form' deformations. In the following, we organize the work based on the garment-related aspects being addressed, such as its role in the avatar, reconstruction, generation, or animation (see Appendix C).

8.1. Avatars with Separate Clothing

Among the existing work on full-body digital avatars, several methods model garments in different layers from the underlying bodies. Such a design choice is possible with different representations, including meshes [XPB*21, XBS*22, XPC*23], NeRF [FYP*22], 3DGS [LLS*24, ZBS*25], or even hybrid representations [ZZY*24]. Avatars with disentangled clothing are inherently more complicated to build compared to those with body and clothing in the same layer, but they offer several key benefits.

First, treating clothing separately alleviates the difficulty of modeling loose-fitting garments such as skirts and dresses [FYP*22, LLS*24]. These garments deviate significantly from the underlying body shape, and are thus hard to represent by just offsets from commonly-used full-body models such as SMPL [LMR*15] and SMPL-X [PCG*19]. Second, this design choice allows the methods to focus on the sophisticated temporal dynamics of the garments beyond skinning techniques that rely primarily on body pose information. Some examples include learning clothing dynamics in the latent space in an autoregressive manner [XPB*21], incorporating physics-based cloth simulation [XBS*22, ZZY*24], or learning clothing deformation from video input to bridge the gap between animation and reconstruction [XPC*23].

8.2. Garment Reconstruction

By garment reconstruction or capture, we refer to methods that convert visual data of garments into 3D animatable assets, including geometry and appearance. Depending on the input, methods can be divided into multi-view reconstruction, which focuses on quality, and monocular reconstruction, which favors applicability. In another dimension, image-based approaches recover static garment assets, and video-based reconstruction focuses on dynamic formation or physical parameters. We distinguish these reconstruction methods from text-conditioned approaches, which we refer to as "generation" of garments and cover mostly in Section 8.3.

8.2.1. Multi-View Reconstruction of Garments

Garment reconstruction usually leverages high-quality surface output from standard multi-view reconstruction algorithms such as Multi-View Stereo (MVS) [FP09] or Photometric Stereo [Woo80]. Early works like Markerless Garment Capture [BPS*08] and Clothcap [PMPHB17, ZPBPM17] focus on segmenting out the garment region from the whole scanned geometry and registering the garment sequence with a template of consistent topology. These segmentation and registration steps turn out to be quite useful for creating the training data for the following works on clothed avatars [XPB*21, XBS*22] or monocular garment reconstruction [BTTPM19]. To further boost the accuracy of registration, some approaches [SSK*05, HSX*22] print color-coded patterns on the fabric to enable reliable feature detection, providing high-quality datasets for learning-based registration approaches [GPX*24].

Some recent work addresses the problem of simulation-ready garment reconstruction, i.e., assets that can be readily consumed by physics-based simulators (including neural simulators). This line of work focuses on two challenges. First, as almost all simulators rely on template meshes for garments, other 3D representations that are used for rendering photorealistic appearances, especially 3DGS, still need to be coupled with meshes for simulation. PhysAvatar [ZZY*24], PGC [GCS*25], and Gaussian Garments [RGW*25] achieve this purpose by embedding Gaussians within the mesh structure, synergizing the physical fidelity of garment dynamics with high-quality appearance from 3DGS.

Second, the dynamic behavior of garments is subject to the influence of physical parameters, including the stretching and bending stiffness of cloth, the sewing pattern or template mesh in the undeformed configuration, as well as the underlying body shape/pose.

Figure 7: *Garment reconstruction methods. Representative images are taken from DressRecon [TXT*25], ChatGarment [BXX*25], HOOD [GBH23] and PhysAvatar [ZZY*24], respectively.*

Compared with work that directly measures forces and deformation in a laboratory setting [WOR11, MBT*12, CTT17, ZLB*24], it is inherently more challenging to estimate these parameters from just visual input, yet more relevant to our discussion on avatars. The basic paradigm is to optimize the estimated parameters by comparing the simulation output with some visual measurement. For example, DiffAvatar [LCL*24] estimates body and clothes parameters through a differentiable eXtended Position-Based Dynamics (XPBD) simulator [SC23]. Physavatar [ZZY*24] utilizes a finite-difference method to compute parameter gradients that optimize a Codimensional-IPC simulation [LKJ21]. CaPhy [SHL*23] and Gaussian Garments [RGW*25], on the other hand, estimate parameters for neural simulators, namely, SNUG [SOC22] and ContourCraft [GBB*24]. Despite such progress, this problem is still challenging. For example, the underlying body pose has a huge influence on the garment dynamics. An optimization that takes into account body pose has a number of Degrees of Freedom (DoF) and requires careful initialization and regularization [LCL*24]. In addition, the simulation of cloth on the human body is contact-rich, for which the gradient estimation can be noisy [ZZY*24].

8.2.2. Monocular Reconstruction of Garments

Accurate reconstruction of separated garments from a single RGB image is a highly challenging problem. Due to the ill-posed nature of this problem, almost all methods in this area rely on some *garment templates*, and learn a neural network that maps single images to the samples or deformation of the templates. For example, Multi-Garment Net [BTTPM19] and BCNet [JZH*20] build PCA subspaces of garments and learn networks to reconstruct PCA coefficients. ClothWild [MNSL22] leverages SMPLicit [CPA*21], a multi-category generative model for implicit garment geometry, to regularize the estimated garment shape. DeClotH [NKOL25] further refines the output geometry and texture from ClothWild [MNSL22] using an SDS optimization [PJBM23]. ReEF [ZQQH22] leverages the surface prediction from a PIFu-style implicit network [SSSJ20], but still registers clothing templates to the predicted surface for high-quality, complete garment shape reconstruction. Garment3DGen [SSX*25] first reconstructs a watertight target geometry using an LRM-style [HZG*24] image-to-3D network, and then deforms a garment template to the target using Jacobian Fields [AGK*22] to obtain high-quality,

simulation-ready mesh assets. Generally speaking, this area is relatively under-explored, possibly due to its inherent difficulty in achieving high-quality output that warrants disentangled predictions and a lack of large-scale, realistic training data.

In addition, several methods have been proposed to reconstruct garments that move separately from the underlying body, given a monocular RGB video as input. These methods usually model the underlying body with SMPL/SMPL-X, but separately track the garment deformation. This problem is less ill-posed than single-image estimation, but still highly challenging when realistic clothing dynamics are desired. MonoClothCap [XPWH20] fits a sequence of clothing shapes in PCA subspaces by differentiable rendering and comparing against the input monocular video. Similar to the idea of monocular dynamic NeRF [PSB*21], several methods [QCZ*23, GJK*24, CPTZ25, TXT*25] simultaneously optimize for a canonical garment shape and its per-frame deformation to perform volumetric differentiable rendering against the input video. In particular, Rec-MV [QCZ*23] and D^3-Human [CPTZ25] address the issue of representing open garment surface, by tracking the boundary curve, and proposing a human manifold Signed Distance Field (hmSDF), respectively. ReLoo [GJK*24] and DressRecon [TXT*25] focus on the challenge of tracking highly dynamic clothing by adopting the bag-of-bones deformation model [PMJ*22, YVN*22].

8.2.3. Sewing Pattern Reconstruction/Generation

A notable trend in garment reconstruction and generation is that, instead of operating in the ambient 3D geometry space, predictions can be made in the 2D space of sewing patterns. Roughly speaking, sewing patterns are 2D templates of garments, defined by outlines and markings, that guide how fabric should be cut and assembled to make the garments. In Computer Graphics (CG) and Computer-Aided Design (CAD), sewing patterns are represented by panels of cloth defined by a set of 2D planar curves, and stitching marks that connect the panels together, forming a compact representation that allows neural networks to make predictions in. For example, NeuralTailor [KL22] proposes to regress sewing patterns from point clouds of the input garments, and Neural Sewing Machine [CWZ*22] trains an autoencoder for sewing patterns that enables monocular garment reconstruction by an additional CNN that maps images to sewing pattern latents. DMap [LCD*25] treats the 2D space of sewing pattern space as a unified UV parameterization of garments, which bridges pixel coordinates and 3D garment geometry for accurate monocular reconstruction.

With the advancement of transformer models, techniques have been developed to tokenize sewing patterns, allowing foundational models to learn their correlation with other modalities such as text and images. Such techniques thus enable reconstruction of sewing patterns from a single image, as in SewFormer [LXL*23], and generation of sewing patterns from text description, as in DressCode [HYZ*24], both through the transformer architecture. AIpparel [NAK*25] and ChatGarment [BXX*25] develop a multimodal foundational model for sewing patterns that can handle multiple garment-related tasks, such as text/image-to-garments and text-based editing. Another advantage of working with sewing patterns is that they can be easily used for cloth simulators as a rest

shape template. For example, Dress-1-to-3 [LYD*25] employs a differentiable cloth simulator to further optimize the sewing patterns predicted by SewFormer [LXL*23], and improves the alignment between the draped garments and the input image.

8.3. Garment Generation

In this section, we discuss recent methods that generate 3D garments, including geometry and texture, from text descriptions. Here, we focus on methods where the generated garments are disentangled from the underlying body. Early methods such as CAPE [MYR*20] and SMPLicit [CPA*21] build generative models (e.g., GANs) that map latent distributions to scanned and registered clothing geometry. As 3D generation methods based on Score Distillation Sampling (SDS) [PJBM23] gain popularity, which optimize 3D representations while requiring only a general 2D image diffusion model, similar techniques have been applied to 3D clothed body generation. Wang and colleagues [WLD*24] generate garment geometry and texture in an offset layer from the underlying SMPL body shape. DAGSM [ZKB*25] generates separate body and garments in two different layers in 2DGS representation, from which garment meshes can be further extracted.

Similar to the trend in general 3D generation that shifts from general 2D-based SDS to 3D-specific priors, recent work starts to leverage garment-specific priors trained on 3D garment data. GarmentDreamer [LLJ*25] and SimAvatar [LYDM*25] train specific diffusion models on garment data for effective generation of base geometry, and then refine the associated 3DGS appearance with SDS. LayerAvatar [ZWLY25] proposes a diffusion model that generates multi-layer body and garment components as 3DGS embedded in disentangled UV regions. Compared with full-body avatars where garments and body are generated in a single layer [KAZ*23], these disentangled avatars enjoy the capability of garment editing and transfer [ZWLY25], as well as realistic physical dynamics from cloth simulators [LLJ*25, LYDM*25].

9. Full-body Avatars

This section traces the evolution of full-body avatar creation along a clear data-efficiency trajectory, beginning with constrained 3D scans, advancing to dynamic multi-view setups, and ultimately to monocular, few-shot, and one-shot methods that enable scalable and personalized digital humans by leveraging powerful generative human priors. Each stage reflects advances in neural modeling and learning-based priors that compensate for reduced input data. This progression highlights the field's shift toward real-world applicability and rapid personalization, ultimately emphasizing the transition from 3D to 2D approaches.

9.1. Full-body Avatars from 3D Scan Data

3D scan-based human avatar modeling aims to reconstruct high-fidelity, animatable digital humans from raw 3D scan data, tackling core challenges such as articulated motion, non-rigid deformations,

and pose generalization, while preserving fine geometric and textural details. Early advances in this domain laid the foundation by integrating neural implicit surfaces with linear blend skinning to enable pose-independent deformation modeling [CZB*21, MYTB21, MZBT21], developing template-free, weakly supervised canonicalization strategies [SYMB21, TSTPM21, QXL*22, LZZ*22], and introducing locally articulated representations to better capture complex clothing dynamics [MSY*21, ZLS*23]. Subsequent work improved physical realism by incorporating physics-based priors into the deformation framework [SHL*23], while others extended full-body modeling to encompass expressive facial and hand articulations through part-aware skinning and pose-conditioned texture networks [SGK*23]. More recently, the focus has shifted toward disentangled and reusable avatar representations to enable controllable reanimation and editing [KKSJ24, HXSH23, BBR*22]. Concurrently, generative frameworks have advanced multi-subject avatar synthesis with stochastic clothing variations [CJS*22], and enabled the direct generation of explicit, textured meshes from 2D normal maps, yielding high-fidelity avatars well-suited for real-time rasterization [ZZC*23].

9.2. Full-body Avatars from Multi-View Video

Moving beyond the requirement of explicit 3D scans, multi-view video enables high-fidelity human avatar reconstruction by providing dense visual observations across both space and time. Recent research has shifted from purely implicit fields toward hybrid and explicit representations, with the aim of improving fidelity, efficiency, and controllability. This section reviews key advances in representation modeling, 3D Gaussian Splatting, and relighting techniques for achieving photorealistic and animatable avatars.

9.2.1. From Implicit Fields to Hybrid Representations

Early approaches to high-fidelity human avatar reconstruction from multi-view video employed encoder-decoder architectures to learn disentangled representations of pose and appearance, enabling view-consistent novel pose synthesis [BWS*21, RBS*22]. These methods laid the foundation for pose-controllable avatar modeling but often suffered from limited view consistency and texture blur under large pose deviations. Neural radiance fields (NeRF) [MST*20] marked a significant shift, enabling high-fidelity view synthesis through implicit volumetric rendering. Applied to human avatars, NeRF-based methods learn continuous radiance functions conditioned on pose, appearance, and spatial coordinates, achieving photorealistic novel view synthesis from multi-view video inputs [PDW*21, PZX*21, ZPX*24, SYZR21], which typically use parametric body models (e.g., SMPL-X [PCG*19]) to define canonical spaces via deformation fields. Despite their visual quality, these implicit representations suffer from high computational cost, slow rendering, and limited geometric controllability. To address these limitations, recent works have transitioned toward hybrid representations that combine explicit geometric priors with neural rendering, moving beyond deformation-driven canonical spaces to structured 3D layouts. These methods tightly integrate parametric models or template meshes with local radiance fields defined on surfaces or volumes, enabling more precise geometric control and efficient rendering. Examples include
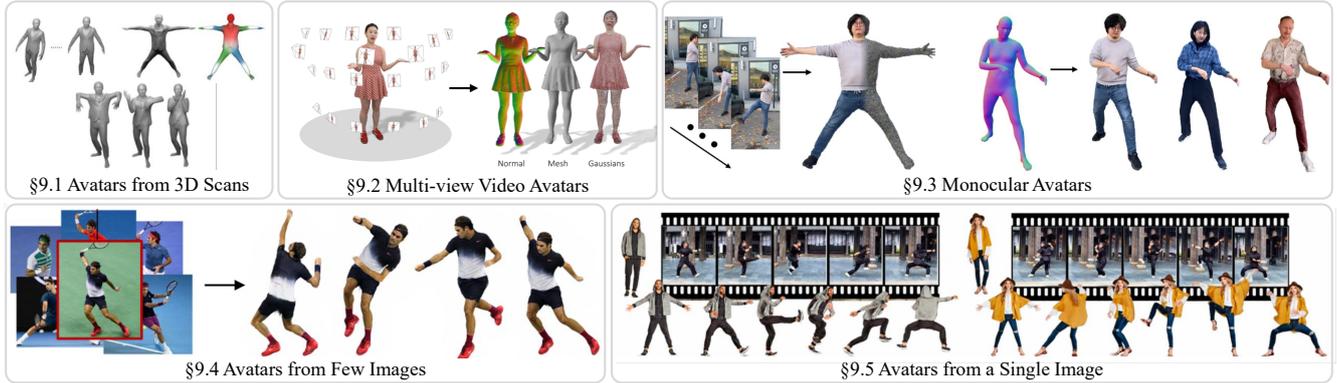
Figure 8: *Overview of 3D full-body avatar research. Representative images are taken from SNARF [CZB*21], TaoAvatar [CHW*25], Vid2Avatar-Pro [GLK*25], PersonNeRF [WSCKS23], and AniGS [QZZ*25].*

surface-aligned radiance volumes [SZZ*22], mesh-anchored light fields [ZHY*22, KLF*23], real-time deep dynamic character models [HLX*21, HLX*23], tri-plane representations for controllable synthesis [ZZTH24], and signed distance function-based reconstructions [CZL*24]. By decoupling geometry and appearance, hybrid schemes also enable efficient texture editing and improved training stability [ZZZ*23, LZL*23]. Temporal coherence is further enhanced via spatio-temporal diffusion models conditioned on pose sequences [JPW*25], marking a shift toward more structured, animatable, and scalable avatar modeling. This evolution from fully implicit to hybrid representations has paved the way for fully explicit and highly efficient frameworks, like 3D Gaussian Splatting.

9.2.2. 3D Gaussian Splatting for Full-body Avatars

3D Gaussian Splatting (3DGS) [KKLD23] has introduced a favorable trade-off between rendering quality, computational efficiency, and controllability, offering faster inference than neural radiance fields with comparable image quality and explicit scene representation. Early efforts adapted 3DGS to animatable human avatars by modeling pose-dependent clothing deformations through template-based 2D parameterization and StyleGAN-driven CNNs [LZWL24], while HuGS [MSD*24] improved visual fidelity via motion-aware Gaussian updates and hierarchical body-part modeling. ASH [PZK*24] further demonstrated efficient and photoreal human rendering through animatable Gaussian splats. Subsequent work has significantly advanced controllability by proposing a two-layer Gaussian representation to disentangle facial expressions from body motion [JSZ*25, ZBS*25], a principle extended by LayGA [LLS*24] to separate clothing from the body for animatable virtual try-on using layered Gaussian representations with geometry-aware supervision. Further enhancing generalization and speed, Zhan *et al.* [ZSYZ25] integrate spatially distributed MLPs with position-based interpolation and learned basis vectors for Gaussian offset prediction, enabling real-time, generalizable avatar rendering. To improve deployment efficiency, SqueezeMe [IPS*25] distills neural pose correctives into lightweight linear layers and shares deformations across neighboring Gaussians, achieving real-time performance for multiple avatars on mobile VR hardware. At the system level, TaoAvatar [CHW*25] enables real-time con-

versational interaction, while Tele-Aloha [TSD*24] demonstrates low-latency streaming of Gaussian-based avatars using lightweight hardware. Meanwhile, extending 3DGS to multi-person scenarios, GeoAvatar [LKL*25] introduces a geometrically-consistent 3DGS framework that preserves spatial relationships and prevents interpenetration through monocular priors and surface ordering. On the physical realism front, PhysAvatar [ZZY*24] integrates inverse physics with mesh-aligned 4D Gaussians and differentiable rendering to estimate fabric material properties from multi-view video, enabling physically plausible cloth dynamics under novel motions. Complementary efforts extend these ideas to sparse-view and egocentric capture setups [SHS*24, SDF*24, CWZ*24, SDZ*25]. While these advances achieve high-fidelity geometry and dynamics, true realism requires consistent appearance under varying illumination. Relighting avatars plausibly in unseen environments is thus essential for their deployment in dynamic VR or AR applications.

9.2.3. Photorealistic Relighting of Full-body Avatars

The evolution of relightable human avatars began with Guo *et al.* [GLD*19], which used a programmable lighting system to capture reflectance maps for high-fidelity image-based relighting. This was followed by learning-based approaches such as Relighting4D [CL22], which introduced a self-supervised neural decomposition of appearance into intrinsic components for relighting from monocular videos. Extensions like IntrinsicAvatar [WAGT24] and RANA [ICN*23] further leverage ray tracing for visibility estimation, promoting disentangled representation of geometry, appearance, and illumination. More recently, methods such as Xu *et al.* [XPG*24] and Carbonera *et al.* [CLGK*24] advanced intrinsic decomposition using 3D-to-UV mapping and distance queries, paving the way for structured representations like Wang *et al.* [WSS*25], which model pose-dependent lighting via zonal harmonics and shadow networks in a canonical Gaussian framework. These are complemented by Li *et al.* [LSZ*23], which use template-guided 2D parameterization for CNN-driven appearance synthesis, and Bolanos *et al.* [BSR24], which enables efficient shadow computation via closed-form occlusion integrals, collectively advancing high-fidelity, generalizable, and efficient relight-

ing. This progression reflects a clear trend toward physically consistent, controllable, and scalable relighting through increasingly explicit modeling of light transport and material appearance.

9.3. Full-body Avatars from Monocular Video

While multi-view setups achieve high fidelity, they remain impractical for most end users. Monocular video provides a far more accessible yet challenging input modality for human avatar creation, demanding high fidelity, efficiency, and robustness under real-world conditions. Recent advances leverage neural and explicit 3D representations to achieve scalable, high-quality reconstruction and animation from casual video. This section reviews progress in appearance modeling and generalization, highlighting the co-evolution of accuracy and practicality.

9.3.1. High-Fidelity Appearance Modeling

Early approaches to monocular human avatar creation employed conditional generative models to synthesize images directly from pose inputs, yielding plausible but often blurry outputs with limited fine detail and temporal stability [LXZ*19, LXH*20]. To overcome these limitations, later methods have embraced advances in neural and hybrid 3D representations, including implicit fields [MST*20] or neural textures [TZN19], enabling view-consistent, photorealistic rendering of high-frequency details [JCSH23, GJC*23, YCW*22, FYP*22, SBR23a, XAS21]. These approaches typically build upon parametric body templates such as SMPL-X [PCG*19] to achieve stable, animatable reconstructions with pose-aware deformations [AMX*18a, AMX*18b, BNT21]. In contrast, template-free methods leverage canonical point clouds to allow greater topological flexibility and more accurate reconstruction of loose clothing and complex garments [SBR23b]. Temporal coherence and motion plausibility are further enhanced through equivariant feature representations and frame prediction strategies [KGE*21, ZWL*22, YCW*22, CGZE19], effectively reducing flickering and rendering artifacts in dynamic sequences.

The integration of 3D Gaussian Splatting has shown significant promise for monocular human avatars by offering high rendering fidelity with low computational overhead, though it faces inherent challenges, including severe depth ambiguity, sparse spatial observations, and temporal flickering under self-occlusion. To address these, recent methods leverage geometric priors from parametric models for robust 3D initialization [HHL24a, HZZ*24], design pose-aware deformation mechanisms via learnable skinning fields or direct mesh-Gaussian coupling, and anchor Gaussians to deformable meshes to enhance geometric accuracy and generalization [SWL*24, PKP*24, WZR*24], thereby achieving high-fidelity, temporally coherent renderings at interactive rates.

To enhance expressiveness, part-aware modeling and dedicated modules for facial and hand articulation enable fine-grained control over local geometry and dynamic details [DZZ*24, HFW*24]. For appearance editing and relighting, IntrinsicAvatar [WAGT24] introduces a Monte Carlo rendering framework for physically accurate decomposition of albedo and illumination, while RANA [ICN*23] enables relightable avatar synthesis from a single video through disentangled modeling of geometry, texture,

and lighting, augmented with synthetic pretraining to improve generalization under real-world conditions.

9.3.2. Efficiency and Generalization to the Wild

Recent advances in monocular human avatar creation have prioritized training efficiency and robust generalization to uncontrolled, real-world environments. A central step has been the development of efficient neural representations that reduce training time from hours to minutes, achieved through strategies such as part-based decomposition with shared priors [GPX*23] or optimized neural rendering pipelines that accelerate convergence without sacrificing visual quality [JCSH23]. Building on this progress, 3D Gaussian Splatting (3DGS) enables real-time, high-fidelity human avatar rendering from monocular video, with improved articulation stability achieved by anchoring Gaussians to deformable template meshes [HHL24a, PKP*24].

To enhance robustness in unconstrained settings, recent methods adopt holistic scene-aware modeling by jointly reconstructing humans and their surroundings using neural scene representations. These approaches disentangle foreground avatars from dynamic backgrounds via dual-field architectures [GJC*23], and extend to multi-person scenarios through layered representations and hybrid 2D/3D supervision [JGK*24]. Moreover, to address the challenge of limited pose coverage in casual video, techniques such as Vid2Avatar-Pro leverage universal priors distilled from large-scale multi-view datasets, enabling high-quality, animatable reconstructions even under sparse or repetitive motion [GLK*25].

9.4. Full-body Avatars from Few-Shot Images

Reducing the input requirement even further, reconstructing animatable human avatars from just a few unconstrained images remains highly under-constrained due to limited pose and appearance variation, prompting recent methods to rely on strong learned priors for generalization. Early approaches use structural and articulatory priors to stabilize reconstruction such as meta-learned initialization of neural SDFs [WMM*21] or keypoint-guided spatial encodings [MBZ*22], thus allowing coherent geometry with few input images and avoiding per-subject fine-tuning. To improve reconstruction completeness under extreme data scarcity, Have-Fun [YCG*24] employs a two-phase optimization that separates identity alignment from appearance completion, enabling robust avatar modeling from very few unconstrained images. For animatability, later works establish canonical spaces by predicting explicit skinning weights for deformation transfer [SLL*24a], or modeling pose-dependent deformations implicitly via shared motion fields conditioned on identity-specific latent codes [WSCKS23]. A new paradigm leverages priors from large-scale multimodal pretraining. PuzzleAvatar [XYL*24] fine-tunes vision-language models to extract disentangled semantic tokens for reconstruction and editing without pose estimation. FRESA [WPW*25] learns a universal generative prior over large-scale 3D human data, enabling zero-shot, feedforward avatar creation with 3D-consistent canonicalization. TAGA [ZCW*25] complements these by learning a self-supervised plausibility prior that corrects implausible skinnings without dense annotations.

9.5. Full-body Avatars from Single Image

Reconstructing a detailed and animatable 3D human avatar from a single image remains a highly ill-posed problem, necessitating strong geometric and appearance priors. The field has evolved along two primary axes: optimization- and regression-based methods that progressively refine 3D geometry through differentiable rendering and structured priors, and emerging generative approaches, particularly diffusion-based models, that synthesize plausible geometry and dynamics by leveraging large-scale data. These paradigms represent complementary strategies for balancing fidelity, generalization, and controllability in avatar creation.

9.5.1. Optimization- or Regression-Based Methods

Early one-shot human avatar reconstructions focused on translating 2D visual inputs into 3D geometry using UV-space representations for mesh reconstruction, advancing beyond basic image-to-image translation tasks [APMTM19, SLH*19, AMB*19]. To improve anatomical accuracy and interpretability, later methods incorporated auxiliary supervision signals, such as anthropometric measurements and semantic attributes [CMH*22], along with spatial layout and depth priors to enhance multi-person pose and shape estimation [SLB*22]. Innovative optimization strategies, like alternating between neural prediction and gradient-based parameter refinement, have enabled high-fidelity reconstructions under limited supervision [KPBD19]. Multi-task learning paradigms also enriched feature semantics by integrating segmentation and keypoint detection tasks [VCR*18]. To address occlusion and ambiguity, part-aware attention mechanisms dynamically reweight visible body regions [KHHB21], while probabilistic approaches model uncertainty in body configurations [ZMZ*23, KPJD21]. The introduction of neural implicit representations marked a significant advancement, enabling high-resolution joint geometry and texture recovery from single images [SHN*19, SSSJ20]. ICON improved pose generalization by conditioning local features on surface normals and refining predictions through a feedback loop [XYTB22], while ECON enhanced structural coherence by fusing normal map predictions with template registration [XYC*23].

9.5.2. Avatar Generation Based on Generative Models

Generative 3D human modeling initially relied on GAN-based frameworks trained on 2D image collections, using compositional NeRFs [HCL*23], articulated tri-plane representations [DCY*23], multi-part models for expressive face and hand control [XZL*23], and unpaired learning for clothed mesh generation [SGY*24]. Subsequent works transitioned to diffusion-based pipelines, operating on structured latent spaces defined on the body manifold [HHL24b] or UV-encoded 3D Gaussians [ZYL*24].

Human video diffusion models. Diffusion models trained on large-scale data have revolutionized one-shot human avatar synthesis, enabling high-fidelity and temporally coherent video generation [BRL*23, BDK*23, ZRA23, CHM*23]. Building on this, UNet-based approaches improve quality and consistency through structured motion conditioning, including 2D skeletons [Hu24], SMPL parameters [ZCD*24], or confidence-aware pose guidance

with progressive latent fusion [ZGW*25]. Further refining controllability, recent works decouple motion representations via token disentanglement [LXS*25] or structure-appearance factorization [WTD*25]. More recently, diffusion transformers (DiTs) have advanced the state-of-the-art by factorizing spatial, temporal, and multi-view dependencies through hierarchical attention, enabling controllable and view-consistent synthesis. Building on this framework, Pippo *et al.* [KWK*25] achieve fully one-shot generation from a single image without relying on parametric models or camera calibration, leveraging large-scale pretraining and attention biasing. To address data scarcity, SIGMAN [YLL*25] introduces HGS-1M, a large-scale 3D Gaussian avatar dataset, and proposes a UV-structured VAE with a DiT-based generator for end-to-end synthesis. IDOL [ZLW*25] directly outputs animatable 3D Gaussians from a single image using a disentangled transformer and synthetic HuGe100K data. Preserving subject identity without sacrificing motion dynamics is addressed through hybrid preference optimization [LJX*25], reward-supervised identity disentanglement [LQZ*25], and multimodal feature fusion [WYL*25].

Concurrently, there is growing interest in scalable, calibration- and pose-free 3D reconstruction. PF-LHM [QLZ*25] enables reconstruction without pose or camera calibration using a Point-Image Transformer and 3D Gaussian Splatting, while PARTE [NKML25] improves texture fidelity in novel views by leveraging 3D part segmentation as a structural prior. To achieve 4D consistency, recent methods incorporate spatiotemporal modeling into diffusion-based frameworks, including interspatial attention [SXS*25], synchronized two-stream optimization [WZN*25], and geometrically guided video diffusion [LDK*25]. To support large viewpoint changes and diverse driving signals, methods correlate source poses with multiple references for detail recovery [HXL*25] and unify audio, video, and multimodal inputs within scalable DiT architectures for rich, controllable animation [LJY*25]. This multimodal paradigm has rapidly expanded through dual-modality transformers for joint video and motion generation [YSC*25], unified world models integrating speech, text, and video [XGL*25, PLT*25], cognitive simulation via large language models for emotionally expressive behavior [JZZ*25], and streaming diffusion architectures for low-latency interactive generation [WZG*26]. Beyond single-character animation, emerging frameworks model interactions through occlusion-aware representations [PSZ*25], region-aware diffusion for contact modeling [LHX*25], environment masking [HWS*25], and depth-guided 3D layout decomposition for interactive editing [MYCB25], multi-identity animation via structural video diffusion [WLZ*25], and affordance-aware scene population [SHM*25], signaling a shift toward physically plausible, context-aware avatars.

Human avatar via auxiliary view synthesis. Recent advances in one-shot 3D avatar generation leverage 2D diffusion models to synthesize auxiliary views or textures from a single in-the-wild image, which are then lifted into 3D for reconstruction. This paradigm effectively mitigates monocular ambiguity while preserving high fidelity and enabling animatability. Early works such as DINAR [SGBL23] and FAMOUS [HAH*24] employ diffusion-based texture synthesis to handle occlusions: DINAR operates in a latent neural texture space, while FAMOUS leverages fashion

datasets to generate plausible back views. AvatarPopUp [KAC*24] decouples 2D view synthesis from 3D reconstruction, achieving multi-view generation in under two seconds and significantly improving scalability. More recent methods extend diffusion-based synthesis to enhance geometric accuracy, identity preservation, and pose controllability. AniGS [QZZ*25] combines transformer-based video diffusion with Gaussian Splatting for dynamic view synthesis and real-time animation. AdaHuman [HYL*25] introduces a pose-conditioned 3D joint diffusion model to improve pose awareness, while PSHuman [LZL*25a] enforces anatomical consistency through cross-scale diffusion under SMPL-X guidance. MoGA [DDS*25] fits a generative 3D Gaussian avatar to synthesized multi-view images by optimizing latent codes, achieving high-fidelity reconstruction. PERSONA [SM25] further improves identity consistency by combining diffusion-based data augmentation with geometry-weighted optimization during fitting.

10. Ethical and Social Implications

As shown in the preceding sections, the state-of-the-art now enables highly realistic digital humans to be created from increasingly sparse inputs. Although these technologies promise broad social and economic benefits, they also raise ethical concerns about potential misuse.

Misinformation. The realism of digital humans allows for powerful forms of misinformation. In particular, AI-generated audiovisual content can convincingly depict people saying or doing things they never did. Such manipulated or created media can be more persuasive than text alone and can spread rapidly, particularly in political or crisis contexts. Early deepfakes already demonstrated misuse risks [TZS*16]; with increased realism, distinguishing genuine from synthetic content has become extremely difficult for human observers.

Trust in visual data. Beyond deliberate misinformation, digital humans and deepfakes pose a broader challenge: they can erode public trust in visual and auditory evidence. Visual and audio media have long been trusted as reliable documentation in journalism, courts, and public discourse. Now that such media can be synthetically generated or altered with high realism, their value as evidence is increasingly questioned. For instance, footage from surveillance cameras may be dismissed as fake even though it is authentic. This growing uncertainty undermines confidence in digital media as a whole, where the mere possibility of manipulation is enough to cause harm even in cases where no manipulation has occurred.

Identity theft. Another potential danger is the ability to impersonate people in online conversations or videos. In particular, *stealing or mimicking* someone's face, voice, or mannerisms can be used to deceive others. A recent real-world incident illustrates this threat: In Hong Kong, fraudsters used a deepfake video call to convincingly impersonate a senior executive and authorize substantial money transfers [gua]. Beyond financial crimes, spoofing biometric identifiers such as facial or voice recognition presents growing security risks as generative tools become more accessible. We expect these challenges to become more prominent as real-time generation methods become more widespread, thus requiring more advanced verification techniques.

Social challenge of remote work / communication. The growing prevalence of remote interaction and avatar-mediated communications poses risks to the quality of human connection. Synthetic avatars could be misused in remote work environments to impersonate colleagues, create fraudulent meetings, or manipulate behavior. The psychological effects include reduced trust, anxiety, or paranoia about whether one is communicating with a real counterpart. There is a risk of increased surveillance or monitoring under the guise of authentication, which may affect worker autonomy.

10.1. Forgery Detection of Synthetic Media

While there are certain challenges, there have also been countermeasures to verify and validate the authenticity of visual content, in particular in the human context.

Active defense. Proactive protection mechanisms embed authenticity at capture time, for example, through secure cameras, cryptographic signatures, or digital watermarking. For instance, the "Secure Digital Camera" concept [BF04] integrates biometric identifiers into captured data, while more recent proposals (e.g., Content Authenticity Initiative) attach provenance metadata directly to files. Although significant progress has been made in watermarking and signature schemes [ZYX*24, NHZ*22], such methods are rarely implemented in consumer hardware.

Method-specific detection. Supervised approaches learn to recognize artifacts of a particular synthesis model. Classic examples include FaceForensics++ [RCV*19] or the Photoshop-based dataset [WWO*19]. Newer works extend this to diffusion-based or neural-rendered content such as DiffusionFake [CYL*24] or DigiFakeAV [LWH*25], which expose generator-specific spectral and temporal patterns. These models achieve high accuracy but generalize poorly to unseen forgery types.

Generalized detection approaches. Manipulation-independent techniques instead assess the physical and statistical plausibility of visual data, testing consistency across lighting, geometry, motion, and temporal coherence. ForensicTransfer [CTR*18] demonstrates how domain-adaptive features can transfer across unseen manipulation methods. ID-Reveal [CRT*20] shifted detection from artifact patterns to identity-based representations, improving robustness to diverse generators. More recently, Zero-Shot Detection [CPNV24] shows that models trained only on real data can identify synthetic content as statistical outliers. While such self-supervised approaches generalize (e.g., to diffusion-based or neural-rendered media), achieving reliability against fully realistic and adversarial content remains an open challenge, underscoring the need for hybrid forensic and provenance-based solutions.

11. Conclusion

This report has reviewed the construction of digital humans through the lens of three key stages: prior learning, personalized avatar creation, and animation. We have reviewed 3D-based avatar methods as well as hybrid 2D-3D models that allow for different types of input controls. We did not cover the generation of such control signals, but gave reference to related reports in this area. In the introduction, we raised the question whether we need 3D avatars in

a world which is increasingly dominated by 2D representations. There is no obvious answer to this and it highly depends on the use case and available data. In the following, we want to further discuss this, in the context of what is missing and how 2D and 3D can help each other.

11.1. What is still missing?

Holistic solutions of creation and animation. Currently, avatar creation and animation are typically decoupled into independent stages. While motion generation can leverage dynamical priors and constraints to enforce physical realism, for example, by ensuring that grasps do not float or that walking motions avoid ground penetration, retargeting these motions to specific 3D avatars introduces new challenges. Discrepancies in body shape and proportions often result in physical inconsistencies such as self-intersections or implausible contact. Recent academic efforts to address these retargeting issues remain limited and typically focus on mesh-based characters in traditional computer graphics pipelines [VCH*21, LKK24]. A promising direction for future work is to explore the co-synthesis of appearance and motion through video generation models [RAFA*25, KNS*25]. Integrating the generation of the avatar's visual appearance with its dynamic behavior within a unified framework could improve physical coherence and reduce artifacts that arise when these stages are treated separately.

Avatar agency. An important aspect of building truly digital humans, beyond motion generation and appearance, is their agency. Unlike traditional animation, there is no external motion signal; the avatars act and behave autonomously, similar to real humans. This means that the avatars can infer the context of a conversation, provide non-verbal cues, and mimic emotions and behaviors. They can interact with and respond to the environment, as well as to other participants within the scene. This capability can be viewed as a step toward the Turing test for avatars. Achieving realistic avatar agency is therefore highly relevant even beyond the digital realm, e.g., in the context of embodied agents [FBC*25]. However, current research efforts in this direction remain limited [AAA*25, KJJ*26, ZZR*25], largely due to data scarcity and the inherent difficulty of modeling and training multimodal avatars. A convincing avatar system must not only observe its environment with a variety of sensors, but also produce a human-like reaction that is appropriate in the social context, all while operating under a strict real-time constraint. In our view, future work on digital humans will increasingly focus on reasoning and agency, as the areas of appearance and motion control continue to reach a higher level of maturity.

General 3D avatar prior. Building truly generative 3D human priors that can be leveraged across various downstream applications, such as avatar inversion from a single or a few input images, is highly challenging, especially when the target reconstruction involves a full 3D avatar. There are several popular approaches:

- **Training on internet-scale datasets**: Leveraging massive and diverse web-scale datasets to improve the generalization and realism of digital humans (e.g., Sapiens [KBM*24]).
- **Using synthetic datasets**: Exploiting high-quality synthetic

datasets to enhance generalization (e.g., David [SAH*25], Syn-Shot [ZGL*25]).

- **Scaling multi-view dataset capture**: Designing systems and pipelines capable of capturing high-quality multi-view data at scale (e.g., NeRSemble [KQG*23], Seamless [BCM*23], DNA-Rendering [CCF*23]).
- **Scaling model capacity to represent everyone**: Expanding model expressivity to accurately capture the full diversity of human appearance, behavior, and motion across populations (e.g., LHM [QGL*25], AniGS [QZZ*25]).

While several works focus on these aspects, one also has to critically ask whether a general avatar prior that can represent any person and asset (hairstyle, garment) can actually be learned. The closest "general" priors that we have today are generative image and video models, as well as large language models that can reason about certain aspects of human behavior. An open question is whether explicitly human-specific priors will remain necessary, given the universality and representational power of foundational models, particularly recent video diffusion systems such as Veo [2,3] (Google), Sora 2 (OpenAI), or Genie [2,3] [PHBB*24], which already model human appearance and motion in a largely domain-agnostic manner.

Hybrid solutions of 2D and 3D avatars. While 2D video generation, exemplified by models such as Sora 2 or Veo 3, demonstrates impressive visual fidelity, it remains limited in controllability, spatiotemporal consistency, and rendering efficiency compared to 3D-based approaches. Recently, adding more fine-grained camera and animation control to video generation models has garnered traction [LXS*25, LXL*25a, CZL*25, LZL*25b]. These endeavors demonstrate the potential of 2D-video based approaches to match existing studio-level 3D avatars in controllability and consistency, while also inheriting useful knowledge about human behavior from their video prior. These approaches lend themselves especially well to more open-ended use-cases such as interactive avatar conversations. Nevertheless, achieving perfect view-consistency with video models at a small real-time budget remains elusive. In the future, 3D avatar priors could serve as conditioning signals or structural guides to enhance 2D video generation. More importantly, 3D digital humans remain indispensable in applications such as VR/AR, the metaverse, 3D virtual try-on, holographic conferencing, and gaming, where unified 3D representations enable multi-user consistency, real-time rendering, and precise geometric fidelity (e.g., accurate garment measurements). Although 3D avatar technology still faces substantial challenges, we anticipate a synergistic evolution of 2D and 3D paradigms, ultimately converging into a unified framework for digital humans.

Spatial awareness. While substantial progress has been made in modeling human–object and human–scene interactions, spatial awareness remains underexplored, particularly in the context of multi-person interaction. Key challenges involve accurate physical simulation, responsiveness to auditory cues, integration of haptic feedback, and even the generation of environments with which agents can meaningfully interact. Consequently, spatial awareness spans not only human behavior modeling, geometry, and appearance, but also spatial cues and the mechanisms by which humans or

virtual agents perceive, interpret, and respond to them. In the long term, advances in spatial awareness may enable avatars to serve as effective learning partners for physical agents.

Avatar benchmarks. Standardized evaluation protocols are essential for fair, transparent, and reproducible assessment of a method's capabilities. However, 3D avatar creation research still lacks widely established benchmarks. Instead, many subfields have converged on de facto evaluation protocols that have emerged organically and are propagated from paper to paper. For instance, evaluations in 3D portrait animation are commonly conducted on the test splits of the VFHQ [XWZ*22] or HDTF [ZLDF21] datasets. In monocular full-body avatar creation, performance is often assessed on selected subjects from ZJU-MoCap [PZX*21] or People-Snapshot [AMX*18b]. Similarly, for 3D hand avatar reconstruction, subsets of the InterHand2.6M dataset [MYW*20] have emerged as the default evaluation setting. However, such loosely defined evaluation protocols always carry the risk of subtle evaluation mismatches that can ultimately render reported numbers incomparable across works. A notable exception is the NeRsemble benchmark [†] which provides a hidden test set and a public leaderboard for two well-defined tasks: Dynamic novel view synthesis of heads and FLAME-driven monocular 3D avatar reconstruction. Overall, however, further efforts are required to unify evaluations across avatar subfields to ensure steady progress. This challenge becomes more pronounced as the capabilities of avatar systems grow. For example, it is less obvious how to evaluate a controllable video generation model that can synthesize avatars from text descriptions and place them into realistically rendered, artificial environments. Developing evaluation protocols that scale with increasingly expressive avatar systems is therefore an open and pressing challenge for the community.

11.2. Beyond 3D Digital Humans

In the future, we hope to see insights from the field of 3D digital humans translated into the physical world. Knowledge about motion and human interaction with objects and scenes could be leveraged to enable humanoid robots to perform analogous actions. Modeling the geometry and physical properties of humans, including the structure of body parts and associated assets (e.g., garments), will play a critical role in facilitating safe and effective human–robot interaction—for instance, when assisting elderly individuals with daily tasks such as dressing. Although most current digital avatar methods do not aim to recover precise physical attributes, such as metrical dimensions or material properties, these capabilities will be essential in the future.

As the field matures across the full pipeline, from representations and priors to per-region modeling and full-body integration, we expect the convergence of appearance, motion, and agency to ultimately yield digital humans that are indistinguishable from, and as capable as, their physical counterparts.

[†] https://kaldir.vc.cit.tum.de/nersemble_benchmark/

Contributor Biographies

Wojciech Zielonka is a Research Scientist at Meta in Pittsburgh. He earned his PhD in 2025 from the Technical University of Darmstadt. Before that, he conducted his doctoral research at the Max Planck Institute for Intelligent Systems (2021–2025). He holds a master's degree in Computer Science from the Technical University of Munich (2018–2021). His research focuses on digital avatars, neural rendering, and human-agent interaction, with particular interests in conversational agents, as well as relighting, hair, and garment modeling.
Webpage: https://zielon.github.io/

Tobias Kirschstein is a PhD student at the Technical University of Munich. He is the builder and maintainer of the NeRSemble dataset, and his research is primarily concerned with photo-realistic 3D head avatars. He received his MSc degree in Computer Science at the Technical University of Munich and was a research scientist intern in the Codec Avatars team of Meta's Reality Lab in Pittsburgh.
Webpage: https://tobias-kirschstein.github.io/

Timo Bolkart is a Senior Research Scientist at Google in Zürich. Previously, he was a Research Scientist at the Max Planck Institute (MPI) for Intelligent Systems in Tübingen (2018–2023), while also a Visiting Academic at Amazon (2021–2023). He joined the MPI as a Postdoctoral Researcher in 2016, after receiving his doctoral degree from Saarland University. During his time at the MPI, he built the widely used parametric head and body models FLAME and SMPL-X. His research centers on digital humans, including the capture and reconstruction of 3D faces and bodies from images and videos, the modeling of their 3D shape, non-rigid deformations, and photorealistic appearance, and their animation with speech.
Webpage: https://sites.google.com/site/bolkartt

Xiang Deng received the B.E. degree from the Institute of Computer and Science, Beijing Jiaotong University, China, in 2020. He is currently the third year Ph.D. student from the Automation Department, Tsinghua University, Beijing, China. His research areas include human motion generation and human video generation.
Webpage: https://scholar.google.com/deng

Vanessa Sklyarova is a Ph.D. student at the Max Planck ETH Center for Learning Systems (CLS), conducting research at the Max Planck Institute for Intelligent Systems and ETH Zurich since 2023. Previously, she was a Research Engineer at Samsung AI Center from 2021 to 2023. She received her M.Sc. in Computer Science from a joint program between the Skolkovo Institute of Science and Technology and the Moscow Institute of Physics and Technology. Her research lies at the intersection of computer vision and computer graphics, focusing on realistic digital humans, neural rendering, and hair reconstruction.
Webpage: https://vanessik.github.io/

Simon Giebenhain is a PhD student at the Technical University of Munich. He is the builder and maintainer of the NPHM dataset, and his research is primarily concerned with 3D reconstruction and tracking of human faces, in order to achieve more lifelike expression control for photo-realistic 3D head avatars. He received his MSc degree in Computer Science from the University of Konstanz.
Webpage: https://simongiebenhain.github.io/
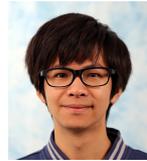
Donglai Xiang is a Research Scientist at NVIDIA Spatial Intelligence Lab (SIL). He obtained his PhD (2019–2023) and MS (2017–2019) degrees from Carnegie Mellon University. His research interest spans computer graphics and computer vision, with a focus on physics-based simulation, geometry processing, dynamic reconstruction and generation. His research work has been published in venues such as SIGGRAPH (Asia), CVPR, ICCV, and received the best paper honorable mention award at 3DV 2020. He contributed to Kaolin and Edify 3D at NVIDIA and Clothed Codec Avatars at Meta.
Webpage: https://xiangdonglai.github.io/

Shunsuke Saito is a Director, Research Scientist at Meta Codec Avatars Lab in Pittsburgh. He obtained his PhD degree at the University of Southern California. His research lies in the intersection of computer graphics, computer vision and machine learning, especially centered around digital human, 3D reconstruction, and performance capture. His work has been published in SIGGRAPH, SIGGRAPH Asia, NeurIPS, ECCV, ICCV and CVPR, three of which have been nominated for CVPR Best Paper Award (2019, 2021) and ECCV Best Paper Award (2024). His real-time volumetric teleportation work also won the Best in Show award at SIGGRAPH 2020 Real-time Live!.
Webpage: https://shunsukesaito.github.io/

Yebin Liu is a Professor at the Department of Automation, Tsinghua University, Beijing, China, where he leads Digital Media Lab. He obtained his PhD degree from Tsinghua University. Prof. Liu's research lies at the intersection of computer vision, computer graphics, and machine learning, where he is particularly interested in 3D reconstruction, neural rendering, human performance capture, and computational photography. His work has been published at leading venues including CVPR, ICCV, ECCV, SIGGRAPH, SIGGRAPH Asia, and TPAMI. He has received multiple awards including the NSFC Excellent Young Scientist Fund and several best paper awards at international conferences.
Webpage: https://www.liuyebin.com/

Matthias Nießner is a Professor at the Technical University of Munich, where he leads the Visual Computing & AI Lab. Before, he was a Visiting Assistant Professor at Stanford University. Prof. Nießner's research lies at the intersection of computer vision, graphics, and machine learning, where he is particularly interested in cutting-edge techniques for 3D generative AI, AI-driven video synthesis, spatial AI, and photorealistic AI Avatars. In total, he has published over 300 academic publications; several of these works won best paper awards, including at SIGCHI'14, HPG'15, SPG'18, and the SIGGRAPH'16 Emerging Technologies Award for the best Live Demo. In addition to his academic impact, Prof. Nießner is a co-founder of Synthesia Inc., a unicorn startup dedicated to democratize synthetic media generation with cutting-edge AI-driven video synthesis technology. Prof. Nießner is also co-founder and CEO of SpAItial AI, a brand-new venture that is at the forefront of generative AI and Spatial Foundation Models.
Webpage: https://niessnerlab.org/publications.html

Justus Thies is a full professor for 3D Graphics & Vision at the Technical University of Darmstadt. He received his doctoral degree from the University of Erlangen-Nuremberg in 2017, was as a postdoc at TU Munich (2017–2021), and a Max Planck Independent Research Group Leader in Tübingen (2021–2024). He has a focus on motion capturing of facial performances, human bodies, and general non-rigid objects, as well as AI-based techniques that allow for photorealistic image and video synthesis. He was awarded with the German Pattern Recognition Award 2024, the Eurographics Young Researcher Award 2024, and an ERC Starting Grant 2024.
Webpage: https://justusthies.github.io/

Appendix A: Physically Based Rendering

In the Computer Graphics field, illumination and material effects are often modeled with the following rendering equation [Kaj86]:

$$\underbrace{L_o(\mathbf{x},\mathbf{v};L_i)}_{\text{Radiance}} = \int_\Omega \underbrace{f(\mathbf{x},\mathbf{l},\mathbf{v})}_{\text{Material}} \underbrace{L_i(\mathbf{x},\mathbf{l})}_{\text{Illumination}} \underbrace{|\mathbf{n}\cdot\mathbf{l}|}_{\text{Cosine law}} d\mathbf{l}. \tag{7}$$

It consists of several parts, which we will discuss in the following. Please note that we explicitly consider the relighting version of Equation 7 and omit the emission term $L_e(\mathbf{x},\mathbf{v})$.

Radiance measures the light energy density traveling in a given direction $\mathbf{v}$. It is the final output of the rendering equation and is perceived by the human visual system as color, based on the three types of cones in the retina [SS99]. In practice, this process is formalized using color matching functions such as CIE 1931 XYZ [WS00], which can then be converted into RGB values.

Material is represented by the BRDF (Bidirectional Reflectance Distribution Function) [oSN77], which describes how light interacts with a surface [PJH23], or BSDF (Bidirectional Scattering Distribution Function), which additionally incorporate transmittance for translucent materials. Intuitively, it tells us how much of the incoming light from a source is reflected or transmitted toward the viewer. It can be modeled using parametric microfacet models like Cook–Torrance [CT82] or neural networks [SRRW21].

Illumination at a point is described by the distribution of incident radiance over all directions. It describes how much light is incoming from the scene and environment, and can be represented by HDR maps [Deb08], a collection of point lights [DHT*00], or Spherical Harmonics [RH01].

Cosine law, also referred to as Lambert's cosine law [Lam60], expresses that the contribution of incoming radiance to a surface point is proportional to the cosine of the angle between the surface normal $\mathbf{n}$ and the incident light direction $\mathbf{l}$, ensuring physically consistent energy weighting.

From Lambertian Material to Relightable Avatars

Over the years, the traditional physically based rendering, as formulated in Equation 7, has been adapted to benefit from neural representations, enhancing quality and simplifying modeling through learning-based approaches in the context of digital humans. In the following sections, we present the evolution from SH
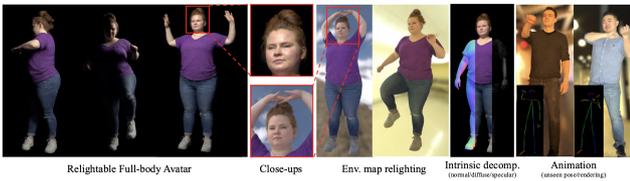

Figure 9: *Relighting requires proper disentangling of material properties such as normals, diffuse, and specular from illumination, often by using OLAT [DHT*00] captures. Image from RFGCA [WSS*25].*

Illumination [RH01,SKS02] to relightable 3DGS avatars [SSS*24, WSS*25,SGN25].

Lambertian surfaces. When assuming a pure Lambertian surface, for which the BRDF is $f_r(\mathbf{x},\mathbf{l},\mathbf{v}) = \rho(\mathbf{x})/\pi$, the rendering equation is view-independent:

$$L_o(\mathbf{x}) = \frac{\rho(x)}{\pi} \int_\Omega L_i(\mathbf{l})\,(n\cdot\mathbf{l})\,d\omega_\mathbf{l}. \tag{8}$$

In the seminal work of Ramamoorthi *et al.* [RH01], the incoming illumination L_i is represented by Spherical Harmonics (SH) [CKK98,Mül06], effectively encoding the irradiance integral into only 9 coefficients. This efficient approximation was quickly adapted to model face illumination by Blanz and Vetter [BV99] and later by Face2Face [TZS*16] for real-time face tracking applications. The equation is denoted as:

$$L_o(\mathbf{x}) \approx \rho(\mathbf{x}) \sum_{l=0}^{2} \sum_{m=-l}^{l} c_{lm} Y_l^m(\mathbf{n}(\mathbf{x})), \tag{9}$$

where the sum is truncated to the first three bands ($l = 0, 1, 2$) of Spherical Harmonics, which were found to be sufficient to approximate diffuse irradiance. In this line of work [BV99,TZS*16, ZBT22], the albedo $\rho(\mathbf{x})$ and geometry are optimizable parameters derived from a statistical face model (3DMM Sec 4.2.2), while the SH coefficients c_{lm} are optimized to represent the scene illumination. The term $\mathbf{n}(\mathbf{x})$ denotes the surface normal at point $\mathbf{x}$, which accounts for the orientation of the face with respect to the incoming light.

View-dependent effects. Kerbl *et al.* [KKLD23] introduced a modified approximation of Eq. 7 that explicitly accounts for the view direction, where each primitive is represented as a small volumetric blob with its own set of SH coefficients:

$$L_o(\mathbf{x},\mathbf{v}) \approx \sum_{l=0}^{L} \sum_{m=-l}^{l} c_{lm} Y_l^m(\mathbf{v}). \tag{10}$$

The major difference compared to Ramamoorthi *et al.* [RH01] is that in 3DGS the spherical harmonics encode the *outgoing radiance* (directional appearance), whereas in Eq. 9 they encode the *incident illumination*. In this sense, 3DGS uses Eq. 10 as a surrogate representation of the full rendering equation (Eq. 7), effectively baking material properties and illumination into the SH coefficients. For further discussion on this interpretation and its deviations from principled volumetric rendering, we refer the reader to the analysis by Celarek *et al.* [CKD*25].

Relighting. The 3DGS [KKLD23] formulation does not support relighting, i.e., the ability to apply arbitrary scene illumination for novel environment rendering. To enable this capability, several modifications to Eq. 10 are required. RGCA [SSS*24] achieve this by explicitly modeling the diffuse component (Eq. 11) and the specular component (Eq. 12) of the rendering equation. In particular, in Eq. 11, novel illumination (e.g., a new HDR environment map denoted as L_i) can be incorporated via the coefficients L_{lm}, which are convolved with the learned radiance transfer functions d_{lm}^k:

$$L_k^{\text{D}} \approx \rho_k \sum_{l=0}^{L_d} \sum_{m=-l}^{l} L_{lm} d_{lm}^k. \tag{11}$$

To properly learn the radiance transfer functions, RGCA requires

one-light-at-a-time (OLAT) captures [DHT*00], which provide controlled illumination and allow disentangling of material properties (ρ_k, d_{lm}^k) from illumination coefficients (L_{lm}). This disentanglement is essential for enabling relighting under novel environments. OLAT can be understood as analogous to PCA [AW10] or Fourier bases [BF*03]: the system probes the face with calibrated white light from individual basis (lights) directions and measures its response. Once these bases are constructed, any novel illumination can be synthesized by projecting an HDR environment map (L_i) onto either linear [Ram02, DHT*00] or neural [LSS*21, WSS*25, LCS*24b, ISS*23] OLAT bases.

Since the diffuse part only models low-frequency lighting effects such as smooth shading, RGCA [SSS*24] incorporates a specular component to capture high-frequency, view-dependent effects in human appearance, such as skin oiliness, eyes, and hair highlights, by using:

$$L_k^S(\mathbf{v}) = v_k(\mathbf{v}) \int_{S^2} L_i(\mathbf{l})\, G_s\big(\mathbf{l}; q_k(\mathbf{v}), \sigma_k\big)\, d\mathbf{l}, \qquad (12)$$

where $G_s(\cdot)$ denotes a spherical Gaussian (SG) kernel [ZLW*21, WRG*09] that compactly approximates specular BRDF lobes and allows a closed-form convolution with environment illumination. Here, $v_k(\mathbf{v})$ denotes the learned view-dependent visibility term, $q_k(\mathbf{v})$ is the reflection direction for Gaussian k and view $\mathbf{v}$, and σ_k controls the sharpness of the specular lobe. In practice, the integral in Eq. 12 is efficiently evaluated for any HDR environment maps by prefiltering them into a mipmap representation [WRG*09], requiring only a single texture lookup per Gaussian [SSS*24]. An alternative design, as proposed in BecomingLit [SGN25], is to use a Cook–Torrance microfacet model [CT82] and split-sum approximation [LDR14] with a two-lobe Blinn–Phong distribution [RGB*20], where only the linear roughness is optimized per Gaussian while the Fresnel and geometry terms remain analytic, simplifying the specular part.

The final neural rendering equation is given by the sum over all Gaussians k:

$$L_o(\mathbf{x}, \mathbf{v}; L_i) \approx \sum_k \alpha_k(\mathbf{x}) \Big[L_k^D + L_k^S(\mathbf{v}) \Big]. \qquad (13)$$

This produces the final radiance as a blend over all face primitives, where α_k is the weight for each Gaussian after volumetric integration. Other methods [LSS*21, LSS*23, WSS*25, LCS*24a, CMG*24] follow a similar approach. However, the major differences arise in how neural components are integrated into the rendering equation, whether based on NeRF, MVP, or 3DGS formulations, and which approximations are used for the diffuse and specular components of Eq. 7. In addition to 3D space, there are several methods that operate only in 2D space, performing single-image relighting [JLL*24, SYH*17, RYCT23, JYG*22, KE19]. In terms of digital humans, these approaches aim to modify the illumination of a portrait or full-body image directly in the image domain, enabling realistic appearance changes without requiring explicit geometry or material reconstruction.

Appendix B: Hair Rendering and Animation

B.1. Rendering and Relighting

Hair rendering is challenging because thousands of fine fibers interact with light through reflection, transmission, and internal scattering, producing glints, highlights, and subtle color shifts. Existing methods can be broadly grouped into physics-based [MJC*03, ZYWK08, ZRL*09, SPJT10, XMR*11, dFH*11, dMH14, CBTB15, KM17, HHH22, YA06, YK08] that describe light transport through and within strands using physical understanding of the world, neural-based approaches [WNS*22, WNS*23, RSW*22, SCD*23, TTS*24, ZSB*24, LOZ*24, ZCW*24b] that try to learn everything from data and hybrid methods [SNA*21, ZCV*25], that utilize physics-based prior for rendering and relighting.

Physics-based rendering. Early approaches such as the Kajiya–Kay model [KK89] treated hair as a collection of oriented cylinders with a simple reflection model. This enabled efficient rendering but could not reproduce physically accurate scattering effects. Physics-based models for hair rendering later advanced through specialized BSDFs that capture light transport in fibers. A major breakthrough came with Marschner *et al.* [MJC*03], who introduced a multi-lobe scattering BSDF for hair fibers. Their model captures surface reflection (R), direct transmission (TT), and transmission with one internal reflection (TRT), which together explain characteristic phenomena such as specular glints, wavelength-dependent color shifts, and the softness of highlights observed in real hair. Building on this foundation, subsequent works improved realism and usability in several ways: Yuksel *et al.* explored the effects of global illumination [YA06] and shadowing from semi-transparent objects [YK08]; Zinke *et al.* [ZYWK08, ZRL*09] introduced approximations for multiple scattering between fibers; Deon *et al.* developed energy-conserving and non-separable fiber scattering models accounting for multiple internal reflections, non-Gaussian longitudinal scattering, azimuthal roughness, caustics, and angle-dependent specular cone shifts [dFH*11, dMH14]; Chiang *et al.* proposed a novel, energy-conserving BFSDF for efficient, expressive, and easily controllable hair and fur [CBTB15]; Khungurn *et al.* studied scattering from elliptical hair fibers, highlighting forward-specular effects from small deviations in cross-section [KM17]; Huang *et al.* introduced the first microfacet-based, non-separable hair BSDF supporting tilted cuticle scales, elliptical cross-sections, and accurate importance sampling while explaining glint-like forward scattering [HHH22]; finally, efforts to combine realism with usability led to art-directable hair shading [SPJT10] and interactive GPU-based rendering using 1D circular Gaussian models for real-time scattering, dynamic editing, and hair eccentricity under complex lighting [XMR*11].

Hair rendering and relighting are implemented in real-time engines, production renderers, and research frameworks. Real-time engines like Unreal Engine [unr] and Unity [Haa14] offer GPU-accelerated strand- or shell-based hair with interactive editing and approximate lighting. Production renderers such as Arnold (Maya) [GIF*18, may], RenderMan [CFS*18], and Blender (Cycles) [Com18] provide physically accurate multi-lobe scattering and full environment illumination. Research frameworks like Mitsuba3 [JSR*22], and NVIDIA OptiX [PBD*10] enable custom BSDFs, differentiable rendering, and multi-scattering experiments.

Neural-based methods. Neural-based approaches learn hair appearance from data by projecting 3D hair geometry into 2D space and supervising through photometric losses. HVH [WNS*22] and NeuWigs [WNS*23] propose to attach unstructured volumetric primitives to hair geometry and use Volumetric Raymarching for rendering. Later, NeuralStrands [RSW*22] show first differentiable rasterization approach for 3D strands, by first projecting them into 2D descriptor map using Line Rasterization and then use rendering network to obtain rgba image. To improve gradient propagation Neural Haircut [SCD*23] and Dr.Hair [TTS*24] instead convert strands into mesh quads and use mesh-based rasterization of quads. Gaussian Haircut [ZSB*24], GaussianHair [LOZ*24], and Groom-Cap [ZCW*24b] propose to use Gaussian Splatting rasterization by attaching gaussians in the middle of each strand segment for differentiable rasterization.

Hybrid methods. Hybrid approaches have emerged that combine prior knowledge from physical hair models with data-driven learning. Sun *et al.* [SNA*21] perform inverse rendering of hair reflectance properties, estimating parameters such as longitudinal roughness and single-strand color. Their method assumes a fixed azimuthal roughness and leverages images captured from multiple viewpoints under varying illumination for optimization. Groom-Light [ZCV*25] extends this idea by inverting the material parameters of a hair BSDF from multi-view OLAT (one-light-at-a-time) images. The optimized parameters include the absorption coefficients, longitudinal and azimuthal roughness, and the interior and exterior indices of refraction. To go beyond the limitations of the analytical model, GroomLight further introduces an implicit light-aware residual representation that captures subtle view- and illumination-dependent appearance details not explained by the parametric BSDF.

B.2. Simulations

Hair motion is a key component in character animation, games, and virtual avatars, but remains challenging due to the large number of strands, complex physical properties, and frequent collisions. Since simulating every strand is computationally prohibitive, most methods simulate a subset of guide hairs and interpolate the rest [LCCZ22]. While interpolation enables fast real-time performance, it often introduces artifacts; to address this, Hsu *et al.* [HWP*24] propose a physics-driven interpolation scheme that leverages guide hair dynamics. Reduced models also struggle with hair-solid interactions, where motion coherence breaks down; Chai *et al.* [CZZ17] alleviates this by adaptively selecting guide hairs and correcting collisions, enabling real-time simulation of over 150K strands with complex hair-solid contacts. Another key challenge is the "sagging effect," where hairstyles lose shape after initialization; Hsu *et al.* [HWP*23] introduce a four-stage sag-free initialization framework that stabilizes quasistatic configurations for strand-based dynamics.

Hair dynamics can be broadly categorized into physics-based simulators [BAC*06, HZS*24, Dav23, HYW*23], data-driven neural approaches [YSZZ19, WNS*22, WNS*23, LXL*25b, ZCP*23], self-supervised methods [SLL*24b] that incorporate physics-based constraints as losses, hybrid approaches [LCCZ22], which use a

physics simulator to generate guiding strands and then interpolate the full hairstyle using neural networks. The following sections provide a detailed description of each category.

Physics-based simulation. Physics-based Simulators explicitly model hair dynamics using physical laws, providing stability and interpretability. Hair is typically represented as strands, rods, or volumes, with different physical models chosen to balance realism and computational efficiency. The most widely used physical models for hair simulation include bone-based simulation, super helices [BAC*06], mass-spring models [HZS*24], and discrete elastic rods [Dav23, HYW*23]. Early hair simulation methods employed a rigid skeleton with bones, where hair strands are approximated by rigid segments attached to a character's mesh and physics constraints drive their motion. This approach is efficient and suitable for real-time applications, though it cannot capture complex strand interactions and requires careful bone design. Therefore, later works introduced more physically accurate models: super helices [BAC*06] represent the curliness and coiling of hair using helical shapes, making wavy and curly hairstyles more realistic; mass-spring models [HZS*24] treat strands as particles connected by springs, balancing simplicity and real-time performance; and discrete elastic rods [Dav23, HYW*23] model strands as elastic rods capturing both bending and twisting, producing high-fidelity hair motion at higher computational cost. Today, physical simulation techniques are widely integrated into *graphics engines* such as Maya [may], Unreal Engine [unr], Blender [Com18], and Digital Salon [HAHZ*24], allowing artists to model, simulate, and render hairstyles with greater realism.

Data-driven methods. Instead of relying on physical priors, several approaches learn hair dynamics directly from data, using monocular video [YSZZ19], multi-view capture [YWO09, XWW*14, WNS*22, WNS*23, LXL*25b], or data generated by a physics simulator [ZCP*23, WXT*25]. These methods require large datasets and often struggle with generalization compared to physics-based simulators. Early approaches rely on classical optimization methods and do not use neural networks. Yamaguchi *et al.* [YWO09] reconstruct realistic dynamic hair from synchronized multi-camera footage with strobe lighting, modeling motion by growing strands segment by segment from root to tip while enforcing orientation consistency and smooth temporal behavior. [XWW*14] uses a motion-path analysis algorithm to track local hair motion across frames, and enforces spatial–temporal coherence through a spacetime optimization framework solved iteratively.

Later, neural network–based approaches were introduced. Yang *et al.* [YSZZ19] capture hair dynamics from monocular video using HairSpatNet to estimate 3D occupancy and orientation and HairTempNet to predict motion across frames. HairSpatNet is trained on both static and dynamic synthetic data, while HairTemp-Net uses only dynamic hair simulated with a mass-spring model. In HVH [WNS*22] hair is represented volumetrically using thousands of primitives and learns dynamics from multi-view data. The method works in two stages: first, guide hairs are tracked with multi-view optical flow and per-frame reconstruction; then, sparse strands are expanded into volumetric primitives, which are optimized with multi-view RGB and optical flow using volumetric

ray marching. NeuWigs [WNS*23] is using differentiable volumetric ray marching to first train an autoencoder to compress hair and head appearance from multi-view RGB images. In a second stage, it introduces a temporal transfer module (T2M) that models hair dynamics in the compressed space, enabling hair animation driven indirectly by head motion and head-relative gravity. Groom-Gen [ZCP*23] learns a quasi-static neural simulator from Houdini data, enabling real-time animation of diverse hairstyles. Driven by head poses, it supports up to 3000 strands running interactively. DGH [WXT*25] further improve hair dynamics driven by head motions, while respecting upper-body collision. Method models dynamic hair appearance to achieve photorealistic novel-view synthesis. In contrast to these methods, HHAvatar [LXL*25b] models head avatars with a layered representation, where hair is represented by unstructured Gaussians. Hair dynamics are captured with an MLP that simulates non-rigid deformations, with motion strength depending on the distance between hair and scalp. The deformation network is conditioned on past hair positions along with current and recent head poses.

Self-supervised methods. Quaffure [SLL*24b] introduces the first real-time quasi-static neural hair simulator trained in a self-supervised manner, eliminating the need for collected data. The method reformulates Cosserat rods for strand representation and leverages physics-based losses—such as elastic potential and penetration—for supervision. At inference, it can animate a wide variety of hairstyles conditioned on body shape and pose.

Hybrid simulators. Hybrid hair simulators leverage prior knowledge of hair physics by simulating a small number of guide strands with a physical simulator and then training a neural interpolator to compute dynamic weights for interpolating the full hairstyle [LCCZ22].

Appendix C: Garment Animation

In this section, we discuss approaches to garment animation, which can be divided into two major categories. One category focuses on producing animation in the 3D space for garment assets that have been either manually created by digital artists, or automatically reconstructed/generated by algorithms discussed in Sec. 8.2 and 8.3. These methods, further divided into physics-based, data-driven, and hybrid methods, are addressed in Sec. C.1. The other category directly generates images and videos of the animated garments in the pixel space, bypassing the process of animating, or even generating, garment assets in 3D, which we discuss in Sec. C.2.

C.1. 3D Garment Animation

Physics-based simulation. Physics-based cloth simulation lays the foundation of garment animation in movies, games, and virtual reality. Modern cloth simulators take form in the seminal work of Baraff and Witkin [BW98], which proposes a formulation based on the implicit Euler integrator, allowing simulations to run in relatively large time steps while remaining stable. Since then, cloth simulation has been extensively studied and improved in all kinds of aspects. For example, the efficiency of cloth simulation benefits from various types of new

solvers, including Position-Based Dynamics (PBD) [MHHR07], XPBD [MMC16], Projective Dynamics [BML*14], and Vertex Block Descent (VBD) [CLYY24]. We have also seen significant improvement in collision handling for cloth in Codimensional-IPC [LKJ21] and followup works [LFL*23, WCL*23], which utilize log-barrier functions together with Continuous Collision Detection (CCD) to ensure penetration-free simulation. A comprehensive review of cloth simulation methods is beyond the scope of this report, which we leave for future work.

Data-driven animation. Physics-based simulation can produce realistic garment motion, but can be slow for accurate, high-resolution results. In addition, it is nontrivial to set up the simulation correctly and keep the simulation stable in the presence of complicated body-cloth collisions. In light of these difficulties, some methods have been proposed to learn clothing deformation with neural networks from simulation results. The basic formulation is to train a network that takes body pose, shape, and a garment template as input, and predicts a garment motion sequence [SOC19, GCS*19, PLPM20, PMJ*22]. To deal with garments of arbitrary 3D topology, convolutional networks on meshes or graphs become a natural choice in these methods [CMM*20, VSGC20]. Some approaches [BMTE21, ZCM22] make predictions in the canonical space to leverage the skinning function, so that the networks only need to learn the residual deformation. LayersNet [SLD23] extends this formulation to handle multi-layer garments. As a powerful generative modeling formulation, diffusion models have also been applied to predict detailed wrinkle on top of base geometry predictions [VGC24] in data-driven cloth animation.

Hybrid neural simulator. A particularly interesting line of work explores leveraging physics as a source of self-supervision for garment animation. Instead of training the neural networks on pre-generated simulated data, some work has demonstrated the possibility of supervising the prediction with physics-inspired losses. These losses have similar or the same definitions as the optimization objectives in the implicit time stepping in cloth simulation, such as strain energy, bending energy, and collision penalty, thus effectively bridging the areas of physical and neural simulators. This line of work starts with the draping of clothes in (quasi-)static scenes [BME21, DLLG*23, KHTG24], and then includes dynamic garment movement as well [SOC22, BME22]. HOOD [GBH23] trains a unified neural network for various garment types, and ContourCraft [GBB*24] handles the collision between multiple layers of garments. EUNet [SLD24] proposes to further learn the constitutive model of the cloth, which defines the mathematical formulation of the strain energy, from an existing cloth trajectory, and then apply this learned energy as part of the physics-inspired loss to train a neural cloth simulator. Interestingly, these methods enjoy both the benefits of physical realism from a physics-informed formulation and the fast inference speed of neural networks.

C.2. Pixel-Space Garment Animation

In this section, we discuss methods that directly generate images or videos of garment animation, which have demonstrated potential for virtual try-on applications. These methods usually assume

a pair of images of a subject and a garment as input, and utilize deep neural networks to generate images of the subject wearing the garment [HWW*18, WZL*18, CPLC21]. Some methods [DLS*19, ALY*21] allows the generated image to be controlled by a target 2D pose or motion sequence, thus truly enabling "animation" of the garments in the synthesized output. While early work often adopts direct supervision or adversarial training [LVKS21], more recent work [ZYZ*23, ZLL*24, LSZ*25] leverages the powerful formulation of diffusion models [HJA20] and DiT [PX23], achieving impressive realism and high resolution output. Beyond static images, generating videos of garments poses a particular challenge, due to the high level of temporal consistency required by garment animation, especially loose-fitting clothing that interacts with the human body and undergoes complicated dynamics [HCW*24, KLL*24, CBVW25, LZY*25]. As video foundation models advance rapidly, it remains an open and intriguing question whether scaling laws alone will enable them to capture garment physics, or whether 3D and physics-based approaches will play a complementary role by offering controllability, consistency, and physical accuracy.

Appendix D: Digital Human Datasets

Table 4 provides an overview of widely used digital human datasets, covering various body parts (face, body, hands, garments, and hair), data types (real, synthetic, or generated), and modalities (images, videos, meshes, or strands). It summarizes dataset scale, camera configurations, and calibration availability, offering a unified reference for benchmarking methods on animatable human modeling and reconstruction.

Dataset	Body Parts	Data Type	Data Modality	Size	# Cameras	Is Calibrated
N NeRSemble [KQG*23]	😊	🖼️	📷	425	16	X
A Ava256 [MKR*24]	😊	🖼️	📷	256	80	X
FS FaceScape [ZYG*23]	😊	🖼️	📷 🕸️	359	60	X
NH NPHM [GKG*23]	😊	🖼️	🕸️	477		
FV FaceVerse [WCY*22]	😊	🖼️	🕸️	128		
M MEAD [WWS*20]	😊	🖼️	📷	60	7	
SR SURREAL [VRM*17]	🏃	🔧	🎥	145	1	
V CelebV-HQ [ZWZ*22]	😊	🖼️	🎥	15653	1	
VC VoxCeleb2 [NCZ17]	😊	🖼️	🎥	6000	1	
V VFHQ [XWZ*22]	😊	🖼️	🎥	7228	1	
F FFHQ [KLA19]	😊	🖼️	📷	70000	1	
AH ActorsHQ [IRG*23]	🏃👕	🖼️	📷	8	160	X
TH THuman2.0 [YZG*21]	🏃	🖼️	🕸️	500	128	X
Z ZJU-MoCap [PZX*21]	🏃	🖼️	📷	9	21	X
HD Human4DiT [SPZ*24]	🏃	🖼️	🕸️	5000	180	X
CD Cloth3D [BME20]	👕	🔧	🕸️	11300		
SP Sewing Pattern [KL21]	👕	🔧	🕸️	23500		
BC BCNet [JZH*20]	👕	🔧	🕸️	54467		
IH InterHand2.6M [MYW*20]	✋	🖼️	📷	27	140	X
H USC-HairSalon [HMLL15]	✂️	🔧	💈	343	-	
CT CT2Hair [SSW*23]	✂️	⚙️	💈	10	-	

Table 4: *Overview of commonly used digital human datasets, detailing covered body parts (face, body, hands, garments, hair), data types (real, synthetic, or generated), and modalities (images, videos, meshes, strands).*

References

[AAA*25] AGRAWAL V., AKINYEMI A., ALVERO K., BEHROOZ M., BUFFALINI J., CARLUCCI F. M., CHEN J., CHEN J., CHEN Z., CHENG S., ET AL.: Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *arXiv preprint arXiv:2506.22554* (2025). 23

[AGK*22] AIGERMAN N., GUPTA K., KIM V. G., CHAUDHURI S., SAITO J., GROUEIX T.: Neural jacobian fields: learning intrinsic mappings of arbitrary meshes. *Transactions on Graphics, (Proc. SIGGRAPH) 41*, 4 (2022), 1–17. 17

[ALY*21] ALBAHAR B., LU J., YANG J., SHU Z., SHECHTMAN E., HUANG J.-B.: Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *Transactions on Graphics (TOG) 40*, 6 (2021), 1–11. 31

[AMB*19] ALLDIECK T., MAGNOR M., BHATNAGAR B. L., THEOBALT C., PONS-MOLL G.: Learning to reconstruct people in clothing from a single rgb camera. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 1175–1186. 21

[AMX*18a] ALLDIECK T., MAGNOR M., XU W., THEOBALT C., PONS-MOLL G.: Detailed human avatars from monocular video. In *3DV* (2018), pp. 98–109. 20

[AMX*18b] ALLDIECK T., MAGNOR M., XU W., THEOBALT C., PONS-MOLL G.: Video based reconstruction of 3d people models. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 8387–8397. 20, 24

[APMTM19] ALLDIECK T., PONS-MOLL G., THEOBALT C., MAGNOR M.: Tex2shape: Detailed full human body geometry from a single image. In *International Conference on Computer Vision (ICCV)* (2019), pp. 2293–2303. 21

[AS00] ASHIKHMIN M., SHIRLEY P.: An anisotropic phong brdf model. *Journal of graphics tools 5*, 2 (2000), 25–32. 6

[ASK*24] ANEJA S., SEVASTOPOLSKY A., KIRSCHSTEIN T., THIES J., DAI A., NIESSNER M.: Gaussianspeech: Audio-driven gaussian avatars, 2024. 6, 8, 9, 11

[ATDN23] ANEJA S., THIES J., DAI A., NIESSNER M.: Clipface: Text-guided editing of textured 3d morphable models. In *SIGGRAPH Conference Papers (SA)* (2023). 11

[AW10] ABDI H., WILLIAMS L. J.: Principal component analysis. *Wiley interdisciplinary reviews: computational statistics 2*, 4 (2010), 433–459. 28

[AYS*24] ABDAL R., YIFAN W., SHI Z., XU Y., PO R., KUANG Z., CHEN Q., YEUNG D.-Y., WETZSTEIN G.: Gaussian shell maps for efficient 3d human generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 9441–9451. 5

[BAC*06] BERTAILS F., AUDOLY B., CANI M.-P., QUERLEUX B., LEROY F., LÉVÊQUE J.-L.: Super-helices for predicting the dynamics of natural hair. *Transactions on Graphics, (Proc. SIGGRAPH) 25*, 3 (2006), 1180–1187. 29

[BBN*12] BEELER T., BICKEL B., NORIS G., BEARDSLEY P., MARSCHNER S., SUMNER R. W., GROSS M.: Coupled 3d reconstruction of sparse facial hair and skin. *Transactions on Graphics, (Proc. SIGGRAPH) 31*, 4 (2012). 14

[BBR*22] BAI Z., BAGAUTDINOV T., ROMERO J., ZOLLHÖFER M., TAN P., SAITO S.: Autoavatar: Autoregressive neural fields for dynamic avatar modeling. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 222–239. 18

[BCM*23] BARRAULT L., CHUNG Y.-A., MEGLIOLI M. C., DALE D., DONG N., DUPPENTHALER M., DUQUENNE P.-A., ELLIS B., ELSA-HAR H., HAAHEIM J., ET AL.: Seamless: Multilingual expressive and streaming speech translation. *arXiv* (2023). 23

[BDK*23] BLATTMANN A., DOCKHORN T., KULAL S., MENDELE-VITCH D., KILIAN M., LORENZ D., LEVI Y., ENGLISH Z., VOLETI V., LETTS A., ET AL.: Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv* (2023). 5, 21

[BF*03] BARON FOURIER J. B. J., ET AL.: *The analytical theory of heat*. Courier Corporation, 2003. 28

[BF04] BLYTHE P., FRIDRICH J.: Secure digital camera. *Digital Investigation* (2004). 22

[BKY*22] BERGMAN A., KELLNHOFER P., YIFAN W., CHAN E., LIN-DELL D., WETZSTEIN G.: Generative neural articulated radiance fields. *Advances in Neural Information Processing Systems (NeurIPS) 35* (2022), 19900–19916. 8, 11

[BME20] BERTICHE H., MADADI M., ESCALERA S.: Cloth3d: clothed 3d humans. In *European Conference on Computer Vision (ECCV)* (2020), Springer, pp. 344–359. 32

[BME21] BERTICHE H., MADADI M., ESCALERA S.: Pbns: physically based neural simulation for unsupervised garment pose space deformation. *Transactions on Graphics, (Proc. SIGGRAPH Asia) 40*, 6 (2021), 1–14. 30

[BME22] BERTICHE H., MADADI M., ESCALERA S.: Neural cloth simulation. *Transactions on Graphics, (Proc. SIGGRAPH) 41*, 6 (2022), 1–14. 30

[BMH*25] BARTHEL F., MORGENSTERN W., HINZER P., HILSMANN A., EISERT P.: Cgs-gan: 3d consistent gaussian splatting gans for high resolution human head synthesis, 2025. 10

[BML*14] BOUAZIZ S., MARTIN S., LIU T., KAVAN L., PAULY M.: Projective dynamics: Fusing constraint projections for fast simulation. *Transactions on Graphics, (Proc. SIGGRAPH) 33*, 4 (2014), 154:1–154:11. 30

[BMTE21] BERTICHE H., MADADI M., TYLSON E., ESCALERA S.: Deepsd: Automatic deep skinning and pose space deformation for 3d garment animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 5471–5480. 30

[BNT21] BUROV A., NIESSNER M., THIES J.: Dynamic surface function networks for clothed human bodies. In *International Conference on Computer Vision (ICCV)* (2021), pp. 10754–10764. 20

[BPS*08] BRADLEY D., POPA T., SHEFFER A., HEIDRICH W., BOUBEKEUR T.: Markerless garment capture. *Transactions on Graphics, (Proc. SIGGRAPH)* (2008), 1–9. 16

[BRL*23] BLATTMANN A., ROMBACH R., LING H., DOCKHORN T., KIM S. W., FIDLER S., KREIS K.: Align your latents: High-resolution video synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 22563–22575. 21

[BSR24] BOLANOS L., SU S.-Y., RHODIN H.: Gaussian shadow casting for neural characters. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 20997–21006. 19

[BTTPM19] BHATNAGAR B. L., TIWARI G., THEOBALT C., PONS-MOLL G.: Multi-garment net: Learning to dress 3d people from images. In *International Conference on Computer Vision (ICCV)* (2019), pp. 5420–5430. 16, 17

[BV99] BLANZ V., VETTER T.: A morphable model for the synthesis of 3d faces. In *Proc. SIGGRAPH* (1999), ACM Press/Addison-Wesley Publishing Co., pp. 187–194. 4, 5, 15, 27

[BW98] BARAFF D., WITKIN A.: Large steps in cloth simulation. In *SIGGRAPH Conference Papers (SA)*. 1998, pp. 43–54. 30

[BWS*21] BAGAUTDINOV T., WU C., SIMON T., PRADA F., SHIRA-TORI T., WEI S.-E., XU W., SHEIKH Y., SARAGIH J.: Driving-signal aware full-body avatars. *Transactions on Graphics, (Proc. SIGGRAPH) 40*, 4 (2021), 1–17. 8, 18

[BXX*25] BIAN S., XU C., XIU Y., GRIGOREV A., LIU Z., LU C., BLACK M. J., FENG Y.: Chatgarment: Garment estimation, generation and editing via large language models. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 2924–2934. 1, 17

[CBC*01] CARR J. C., BEATSON R. K., CHERRIE J. B., MITCHELL T. J., FRIGHT W. R., MCCALLUM B. C., EVANS T. R.: Reconstruction and representation of 3d objects with radial basis functions. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* (2001), pp. 67–76. 4

[CBTB15] CHIANG M. J.-Y., BITTERLI B., TAPPAN C., BURLEY B.: A practical and controllable hair and fur model for production path tracing. In *SIGGRAPH Conference Papers (SA)* (2015). 28

[CBVW25] CHEN J.-K., BANSAL A., VO M. P., WANG Y.-X.: Virtual fitting room: Generating arbitrarily long videos of virtual try-on from a single image. In *Advances in Neural Information Processing Systems (NeurIPS)* (2025). 31

[CCF*23] CHENG W., CHEN R., FAN S., YIN W., CHEN K., CAI Z., WANG J., GAO Y., YU Z., LIN Z., ET AL.: Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *International Conference on Computer Vision (ICCV)* (2023), pp. 19982–19993. 23

[CCH*25] CHEN Y., CARRASCO F. V., HÄNE C., NAM G., BAZIN J.-C., DE LA TORRE F.: Doubly hierarchical geometric representations for strand-based human hairstyle generation. In *Advances in Neural Information Processing Systems (NeurIPS)* (2025). 12, 14

[CDA*24] CHHATRE K., DANĚČEK R., ATHANASIOU N., BECHERINI G., PETERS C., BLACK M. J., BOLKART T.: AMUSE: Emotional speech-driven 3D body animation via disentangled latent diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 1942–1953. 6

[CFS*18] CHRISTENSEN P., FONG J., SHADE J., WOOTEN W., SCHUBERT B., KENSLER A., FRIEDMAN S., KILPATRICK C., RAMSHAW C., BANNISTER M., RAYNER B., BROUILLAT J., LIANI M.: Renderman: An advanced path-tracing architecture for movie rendering. *Transactions on Graphics (TOG) 37*, 3 (2018). 28

[CGH*25] CHENG G., GAO X., HU L., HU S., HUANG M., JI C., LI J., MENG D., QI J., QIAO P., SHEN Z., SONG Y., SUN K., TIAN L., WANG F., WANG G., WANG Q., WANG Z., XIAO J., XU S., ZHANG B., ZHANG P., ZHANG X., ZHANG Z., ZHOU J., ZHUO L.: Wananimate: Unified character animation and replacement with holistic replication, 2025. 5, 11

[CGZE19] CHAN C., GINOSAR S., ZHOU T., EFROS A. A.: Everybody dance now. In *International Conference on Computer Vision (ICCV)* (2019), pp. 5933–5942. 20

[CH24] CHU X., HARADA T.: Generalizable and animatable gaussian head avatar. *Advances in Neural Information Processing Systems (NeurIPS) 37* (2024), 57642–57670. 8, 10

[CHM*23] CHEN Z., HONG F., MEI H., WANG G., YANG L., LIU Z.: Primdiffusion: Volumetric primitives diffusion for 3d human generation. *Advances in Neural Information Processing Systems (NeurIPS) 36* (2023), 13664–13677. 21

[CHV*22] CORONA E., HODAN T., VO M., MORENO-NOGUER F., SWEENEY C., NEWCOMBE R., MA L.: Lisa: Learning implicit shape and appearance of hands. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 20533–20543. 6, 15

[CHW*25] CHEN J., HU J., WANG G., JIANG Z., ZHOU T., CHEN Z., LV C.: Taoavatar: Real-time lifelike full-body talking avatars for augmented reality via 3d gaussian splatting. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025). 1, 3, 8, 19

[CJS*22] CHEN X., JIANG T., SONG J., YANG J., BLACK M. J., GEIGER A., HILLIGES O.: Gdna: Towards generative detailed neural avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 20427–20437. 18

[CKD*25] CELAREK A., KOPANAS G., DRETTAKIS G., WIMMER M., KERBL B.: Does 3d gaussian splatting need accurate volumetric rendering? In *Computer Graphics Forum* (2025), Wiley Online Library, p. e70032. 27

[CKK98] COLTON D. L., KRESS R., KRESS R.: *Inverse acoustic and electromagnetic scattering theory*, vol. 93. Springer, 1998. 27

[CL22] CHEN Z., LIU Z.: Relighting4d: Neural relightable human from videos. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 606–623. 19

[CLC*22] CHAN E. R., LIN C. Z., CHAN M. A., NAGANO K., PAN B., DE MELLO S., GALLO O., GUIBAS L. J., TREMBLAY J., KHAMIS S., ET AL.: Efficient geometry-aware 3d generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), Computer Vision Foundation / IEEE, pp. 16123–16133. 5, 9, 10, 11

[CLGK*24] CARBONERA LUVIZON D., GOLYANIK V., KORTYLEWSKI A., HABERMANN M., THEOBALT C.: Relightable neural actor with intrinsic decomposition and pose control. In *European Conference on Computer Vision (ECCV)* (2024), pp. 465–483. 19

[CLS*15] CHAI M., LUO L., SUNKAVALLI K., CARR N., HADAP S., ZHOU K.: High-quality hair modeling from a single portrait photo. *Transactions on Graphics, (Proc. SIGGRAPH) 34*, 6 (2015). 13

[CLYY24] CHEN A. H., LIU Z., YANG Y., YUKSEL C.: Vertex block descent. *Transactions on Graphics, (Proc. SIGGRAPH) 43*, 4 (2024), 1–16. 30

[CLZ*25] CHEN Y., LIANG S., ZHOU Z., HUANG Z., MA Y., TANG J., LIN Q., ZHOU Y., LU Q.: Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters, 2025. 11

[CMG*24] CHEN Z., MOON G., GUO K., CAO C., PIDHORSKYI S., SIMON T., JOSHI R., DONG Y., XU Y., PIRES B., ET AL.: Urhand: Universal relightable hands. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 119–129. 8, 15, 16, 28

[CMH*22] CHOUTAS V., MÜLLER L., HUANG C.-H. P., TANG S., TZIONAS D., BLACK M. J.: Accurate 3d body shape regression using metric and semantic attributes. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 2718–2728. 21

[CMM*20] CHENTANEZ N., MACKLIN M., M"ULLER M., JESCHKE S., KIM T.-Y.: Cloth and skin deformation with a triangle mesh based convolutional neural network. *Computer Graphics Forum (CGF) 39*, 8 (2020), 123–134. 30

[CMT*25] CHEN K., MOHAN S., THEISS J., OPREA S., SRIDHAR S., PRAKASH A.: Interactavatar: Modeling hand-face interaction in photorealistic avatars with deformable gaussians. In *International Conference on Computer Vision (ICCV)* (2025), pp. 10410–10420. 16

[CMZ*25] CHEN K., MIN C., ZHANG L., HAMPALI S., KESKIN C., SRIDHAR S.: Foundhand: Large-scale domain-specific learning for controllable hand image generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 17448–17460. 16

[Com18] COMMUNITY B. O.: Blender - a 3d modelling and rendering package, 2018. 11, 28, 29

[CPA*21] CORONA E., PUMAROLA A., ALENYA G., PONS-MOLL G., MORENO-NOGUER F.: Smplicit: Topology-aware generative model for clothed people. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 11875–11885. 17, 18

[CPLC21] CHOI S., PARK S., LEE M., CHOO J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 14131–14140. 31

[CPNV24] COZZOLINO D., POGGI G., NIEßNER M., VERDOLIVA L.: Zero-shot detection of ai-generated images, 2024. 22

[CPTZ25] CHEN H., PENG B., TAO Y., ZHANG J.: D^3-human: Dynamic disentangled digital human from monocular video. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 10836–10846. 17

[CRT*20] COZZOLINO D., RÖSSLER A., THIES J., NIESSNER M., VERDOLIVA L.: Id-reveal: Identity-aware deepfake video detection. *arXiv* (2020). 22

[CSK*22] CAO C., SIMON T., KIM J. K., SCHWARTZ G., ZOLLHOEFER M., SAITO S., LOMBARDI S., WEI S.-E., BELKO D., YU S.-I., ET AL.: Authentic volumetric avatars from a phone scan. *Transactions on Graphics, (Proc. SIGGRAPH) 41*, 4 (2022), 1–19. 9

[CSW*16] CHAI M., SHAO T., WU H., WENG Y., ZHOU K.: Autohair: Fully automatic hair modeling from a single image. *Transactions on Graphics, (Proc. SIGGRAPH) 35* (2016), 1–12. 13

[CT82] COOK R. L., TORRANCE K. E.: A reflectance model for computer graphics. *Transactions on Graphics, (Proc. SIGGRAPH) 1*, 1 (1982), 7–24. 27, 28

[CTR*18] COZZOLINO D., THIES J., RÖSSLER A., RIESS C., NIESSNER M., VERDOLIVA L.: Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv* (2018). 22

[CTT17] CLYDE D., TERAN J., TAMSTORF R.: Modeling and data-driven parameter estimation for woven fabrics. In *Symposium on Computer Animation (SCA)* (2017), pp. 1–11. 17

[CWS23] CHEN X., WANG B., SHUM H.-Y.: Hand avatar: Free-pose hand animation and rendering from monocular video. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 8683–8693. 8, 15

[CWW*12] CHAI M., WANG L., WENG Y., YU Y., GUO B., ZHOU K.: Single-view hair modeling for portrait manipulation. *Transactions on Graphics, (Proc. SIGGRAPH) 31*, 4 (2012). 13

[CWW*13] CHAI M., WANG L., WENG Y., JIN X., ZHOU K.: Dynamic hair manipulation in images and videos. *Transactions on Graphics, (Proc. SIGGRAPH) 32*, 4 (2013). 12, 13

[CWZ*22] CHEN X., WANG G., ZHU D., LIANG X., TORR P., LIN L.: Structure-preserving 3d garment modeling with neural sewing machines. *Advances in Neural Information Processing Systems (NeurIPS) 35* (2022), 15147–15159. 17

[CWZ*24] CHEN J., WANG J., ZHANG Y., PANDEY R., BEELER T., HABERMANN M., THEOBALT C.: EgoAvatar: Egocentric view-driven and photorealistic full-body avatars. In *SIGGRAPH Asia Conference Papers (SA)* (2024), pp. 1–11. 19

[CYL*24] CHEN S., YAO T., LIU H., SUN X., DING S., JI R., ET AL.: Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. *Advances in Neural Information Processing Systems 37* (2024), 101474–101497. 22

[CZB*21] CHEN X., ZHENG Z., BLACK M. J., HILLIGES O., GEIGER A.: Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)* (2021), pp. 11594–11604. 18, 19

[CZL*24] CHEN Y., ZHENG Z., LI Z., XU C., LIU Y.: Meshavatar: Learning high-quality triangular human avatars from multi-view videos, 2024. 19

[CZL*25] CAO C., ZHOU J., LI S., LIANG J., YU C., WANG F., XUE X., FU Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers* (2025), pp. 1–12. 5, 23

[CZZ17] CHAI M., ZHENG C., ZHOU K.: Adaptive skinning for interactive hair-solid simulation. *Transactions on Visualization and Computer Graphics (TVCG) 23*, 7 (2017), 1725–1738. 29

[Dav23] DAVIET G.: Interactive hair simulation on the gpu using admm. In *SIGGRAPH Conference Papers (SA)* (New York, NY, USA, 2023), SIGGRAPH '23, Association for Computing Machinery. 29

[DCK*22] DROBYSHEV N., CHELISHEV J., KHAKHULIN T., IVAKHNENKO A., LEMPITSKY V., ZAKHAROV E.: Megaportraits: One-shot megapixel neural head avatars. In *International Conference on Multimedia and Expo (ICME)* (2022), pp. 2663–2671. 11

[DCT*23] DANĚČEK R., CHHATRE K., TRIPATHI S., WEN Y., BLACK M., BOLKART T.: Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia Conference Papers (SA)* (2023), ACM. 6

[DCY*23] DONG Z., CHEN X., YANG J., BLACK M. J., HILLIGES O., GEIGER A.: AG3D: Learning to generate 3D avatars from 2D image collections. In *International Conference on Computer Vision (ICCV)* (2023), IEEE, pp. 14870–14881. 21

[DDS*25] DONG Z., DUAN L., SONG J., BLACK M. J., GEIGER A.: Moga: 3d generative avatar prior for monocular gaussian avatar reconstruction. In *International Conference on Computer Vision (ICCV)* (2025), pp. 1–11. 22

[Deb08] DEBEVEC P.: Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Transactions on Graphics, (Proc. SIGGRAPH)*. ACM, 2008, pp. 189–198. 27

[dFH*11] D'EON E., FRANCOIS G., HILL M., LETTERI J., AUBRY J.-M.: An energy-conserving hair reflectance model. In *Computer Graphics Forum (CGF)* (2011), pp. 1181–1187. 28

[DHT*00] DEBEVEC P., HAWKINS T., TCHOU C., DUIKER H.-P., SAROKIN W., SAGAR M.: Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques* (2000), pp. 145–156. 6, 27, 28

[dLGHT*25] DE LA GORCE M., HEWITT C., TAKACS T., GERDISCH R., HOSENIE Z., MEISHVILI G., KOWALSKI M., CASHMAN T. J., CRIMINISI A.: VoluMe – authentic 3d video calls from live gaussian splat prediction. In *International Conference on Computer Vision (ICCV)* (2025). 6, 10

[DLLG*23] DE LUIGI L., LI R., GUILLARD B., SALZMANN M., FUA P.: Drapenet: Garment generation and self-supervised draping. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 1451–1460. 30

[DLS*19] DONG H., LIANG X., SHEN X., WANG B., LAI H., ZHU J., HU Z., YIN J.: Towards multi-pose guided virtual try-on network. In *International Conference on Computer Vision (ICCV)* (2019), pp. 9026–9035. 31

[dMH14] D'EON E., MARSCHNER S., HANIKA J.: A fiber scattering model with non-separable lobes. In *SIGGRAPH Conference Papers (SA)* (2014). 28

[DWR*24] DENG Y., WANG D., REN X., CHEN X., WANG B.: Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 7119–7130. 5

[DZX*23] DING Z., ZHANG X., XIA Z., JEBE L., TU Z., ZHANG X.: Diffusionrig: Learning personalized priors for facial appearance editing. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 12736–12746. 5, 11

[DZZ*24] DENG X., ZHENG Z., ZHANG Y., SUN J., XU C., YANG X., WANG L., LIU Y.: Ram-avatar: Real-time photo-realistic avatar from monocular videos with full-body control. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 1996–2007. 20

[EST*20] EGGER B., SMITH W. A., TEWARI A., WUHRER S., ZOLLHOEFER M., BEELER T., BERNARD F., BOLKART T., KORTYLEWSKI A., ROMDHANI S., ET AL.: 3d morphable face models—past, present, and future. *Transactions on Graphics, (Proc. SIGGRAPH) 39*, 5 (2020), 1–38. 2, 5

[FBC*25] FUNG P., BACHRACH Y., CELIKYILMAZ A., CHAUDHURI K., CHEN D., CHUNG W., DUPOUX E., GONG H., JÉGOU H., LAZARIC A., ET AL.: Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355* (2025). 23

[FCB*21] FENG Y., CHOUTAS V., BOLKART T., TZIONAS D., BLACK M.: Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)* (2021), pp. 792–804. 6

[FFBB21] 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*, 4 (2021), 1–13. 6

[FP09] FURUKAWA Y., PONCE J.: Accurate, dense, and robust multiview stereopsis. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 32*, 8 (2009), 1362–1376. 16

[FPK*24] FAN Z., PARELLI M., KADOGLOU M. E., CHEN X., KOCABAS M., BLACK M. J., HILLIGES O.: Hold: Category-agnostic 3d

reconstruction of interacting hands and objects from video. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 494–504. 1, 15, 16

[FYA*24] Fu Q., Yang X., Asad M., Oh C., Yuan S., Slabaugh G.: Adaptive multi-modal control of digital human hand synthesis using a region-aware cycle loss. In *European Conference on Computer Vision (ECCV)* (2024), Springer, pp. 68–84. 16

[FYP*22] Feng Y., Yang J., Pollefeys M., Black M. J., Bolkart T.: Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia Conference Papers (SA)* (2022), pp. 1–9. 6, 12, 16, 20

[GBB*24] Grigorev A., Becherini G., Black M., Hilliges O., Thomaszewski B.: Contourcraft: Learning to resolve intersections in neural multi-garment simulations. In *SIGGRAPH Conference Papers (SA)* (2024), pp. 1–10. 17, 30

[GBH23] Grigorev A., Black M. J., Hilliges O.: Hood: Hierarchical graphs for generalized modelling of clothing dynamics. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 16965–16974. 1, 17, 30

[GCS*19] Gundogdu E., Constantin V., Seifoddini A., Dang M., Salzmann M., Fua P.: Garnet: A two-stream network for fast and accurate 3d cloth draping. In *International Conference on Computer Vision (ICCV)* (2019), pp. 8739–8748. 30

[GCS*25] Guo M., Chiang M. J.-Y., Santesteban I., Sarafianos N., Chen H.-y., Halimi O., Božič A., Saito S., Wu J., Liu C. K., et al.: Pgc: Physics-based gaussian cloth from a single pose. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025). 16

[GIF*18] Georgiev I., Ize T., Farnsworth M., Montoya-Vozmediano R., King A., Lommel B. V., Jimenez A., Anson O., Ogaki S., Johnston E., Herubel A., Russell D., Servant F., Fajardo M.: Arnold: A brute-force production path tracer. *Transactions on Graphics (TOG) 37*, 3 (2018). 28

[GJC*23] Guo C., Jiang T., Chen X., Song J., Hilliges O.: Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 12858–12868. 20

[GJK*24] Guo C., Jiang T., Kaufmann M., Zheng C., Valentin J., Song J., Hilliges O.: Reloo: Reconstructing humans dressed in loose garments from monocular video in the wild. In *European Conference on Computer Vision (ECCV)* (2024), pp. 21–38. 17

[GKG*23] Giebenhain S., Kirschstein T., Georgopoulos M., Rünz M., Agapito L., Niessner M.: Learning neural parametric head models. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 3, 5, 32

[GKG*24] Giebenhain S., Kirschstein T., Georgopoulos M., Rünz M., Agapito L., Niessner M.: Mononphm: Dynamic head reconstruction from monocular videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 10747–10758. 5, 9

[GKR*24] Giebenhain S., Kirschstein T., Rünz M., Agapito L., Niessner M.: Npga: Neural parametric gaussian avatars. In *SIGGRAPH Asia Conference Papers (SA)* (2024), pp. 1–11. 9

[GLD*19] Guo K., Lincoln P., Davidson P., Busch J., Yu X., Whalen M., Harvey G., Orts-Escolano S., Pandey R., Dourgarian J., et al.: The relightables: Volumetric performance capture of humans with realistic relighting. *Transactions on Graphics (TOG) 38*, 6 (2019), 1–19. 19

[GLK*25] Guo C., Li J., Kant Y., Sheikh Y., Saito S., Cao C.: Vid2avatar-pro: Authentic avatar from videos in the wild via universal prior. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025). 1, 8, 19, 20

[GLW*25] Gu M., Li J., Wu Y., Luo H., Zheng J., Bai X.: 3d human avatar reconstruction with neural fields: A recent survey. *Image Vision Comput. 154*, C (2025). 2

[GPL*22] Grassal P.-W., Prinzler M., Leistner T., Rother C., Niessner M., Thies J.: Neural head avatars from monocular rgb videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 18653–18664. 8, 9

[GPX*23] Geng C., Peng S., Xu Z., Bao H., Zhou X.: Learning neural volumetric representations of dynamic humans in minutes. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 8759–8770. 20

[GPX*24] Guo J., Prada F., Xiang D., Romero J., Wu C., Park H. S., Shiratori T., Saito S.: Diffusion shape prior for wrinkle-accurate cloth registration. In *International Conference on 3D Vision (3DV)* (2024), IEEE, pp. 790–799. 16

[GTZN21] Gafni G., Thies J., Zollhofer M., Niessner M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 8649–8658. 6, 8, 9

[gua] The guardian – deepfake video conference. Deepfake video conference. Accessed: 2025-10-10. 22

[GZL*24] Guo J., Zhang D., Liu X., Zhong Z., Zhang Y., Wan P., Zhang D.: Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv* (2024). 11

[GZL*25] Gao X., Zhou J., Liu D., Zhou Y., Zhang J.: Controlling avatar diffusion with learnable gaussian embedding, 2025. 11

[GZW*23] Guo Z., Zhou W., Wang M., Li L., Li H.: Handnerf: Neural radiance fields for animatable interacting hands. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 21078–21087. 15

[Haa14] Haas J. K.: A history of the unity game engine, 2014. 28

[HAH*24] Hema V. M., Aich S., Haene C., Bazin J.-C., De la Torre F.: Famous: High-fidelity monocular 3d human digitization using view synthesis. In *European Conference on Computer Vision (ECCV)* (2024), pp. 57–73. 21

[HAHZ*24] He C., Amador Herrera J. A., Zhou Y., Shu Z., Sun X., Feng Y., Pirk S., Michels D. L., Zhang M., Wang T. Y., Rushmeier H.: Digital salon: An ai and physics-driven tool for 3d hair grooming and simulation. In *SIGGRAPH Asia 2024 Real-Time Live!* (2024), SA Real-Time Live! '24. 29

[HBLB17] Hu L., Bradley D., Li H., Beeler T.: Simulation-ready hair capture. *Computer Graphics Forum (CGF)* (2017), 281–294. 13

[HCB*24] HaCohen Y., Chiprut N., Brazowski B., Shalem D., Moshe D., Richardson E., Levin E., Shiran G., Zabari N., Gordon O., et al.: Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103* (2024). 5

[HCL*23] Hong F., Chen Z., Lan Y., Pan L., Liu Z.: EVA3D: Compositional 3D human generation from 2D image collections. In *International Conference on Learning Representations (ICLR)* (2023), OpenReview.net. 21

[HCT*24] He M., Clausen P., Taşel A. L., Ma L., Pilarski O., Xian W., Rikker L., Yu X., Burgert R., Yu N., Debevec P.: Diffrelight: Diffusion-based facial performance relighting. In *SIGGRAPH Asia Conference Papers (SA)* (2024). 9

[HCW*24] He Z., Chen P., Wang G., Li G., Torr P. H., Lin L.: Wildvidfit: Video virtual try-on in the wild via image-based controlled diffusion models. In *European Conference on Computer Vision (ECCV)* (2024), Springer, pp. 123–139. 31

[HDZ*22] Hong W., Ding M., Zheng W., Liu X., Tang J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868* (2022). 5

[HFW*24] Hu H., Fan Z., Wu T., Xi Y., Lee S., Pavlakos G., Wang Z., et al.: Expressive gaussian human avatars from monocular rgb video. *Advances in Neural Information Processing Systems (NeurIPS) 37* (2024), 5646–5660. 20

[HGY*25] HE Y., GU X., YE X., XU C., ZHAO Z., DONG Y., YUAN W., DONG Z., BO L.: Lam: Large avatar model for one-shot animatable gaussian head. In *SIGGRAPH Conference Papers (SA)* (2025), pp. 1–13. 10

[HGZ*23] HU L., GAO X., ZHANG P., SUN K., ZHANG B., BO L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117* (2023). 2, 5

[HHH22] HUANG W., HULLIN M., HANIKA J.: A microfacet-based hair scattering model. *Computer Graphics Forum (CGF) 41* (2022), 79–91. 28

[HHL24a] HU S., HU T., LIU Z.: Gauhuman: Articulated gaussian splatting from monocular human videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 20418–20431. 20

[HHL24b] HU T., HONG F., LIU Z.: StructLDM: Structured latent diffusion for 3D human generation. In *European Conference on Computer Vision (ECCV)* (2024), vol. 15109, Springer, pp. 363–381. 21

[HJA20] HO J., JAIN A., ABBEEL P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems 33* (2020), 6840–6851. 31

[HK93] HANRAHAN P., KRUEGER W.: Reflection from layered surfaces due to subsurface scattering. In *SIGGRAPH*. ACM, 1993, pp. 165–174. 6

[HLK*25] HE C., LI J., KIRSCHSTEIN T., SEVASTOPOLSKY A., SAITO S., TAN Q., ROMERO J., CAO C., RUSHMEIER H., NAM G.: 3dgh: 3d head generation with composable hair and face. *Transactions on Graphics, (Proc. SIGGRAPH) 44*, 4 (2025), 1–12. 11, 12, 14

[HLL*24] HUANG X., LI H., LIU W., LIANG X., YAN Y., CHENG Y., GAO C.: Learning interaction-aware 3d gaussian splatting for one-shot hand avatars. *Advances in Neural Information Processing Systems 37* (2024), 14127–14147. 15

[HLX*21] HABERMANN M., LIU L., XU W., ZOLLHOEFER M., PONS-MOLL G., THEOBALT C.: Real-time deep dynamic characters. *Transactions on Graphics, (Proc. SIGGRAPH) 40*, 4 (2021), 1–16. 19

[HLX*23] HABERMANN M., LIU L., XU W., PONS-MOLL G., ZOLL-HOEFER M., THEOBALT C.: Hdhumans: A hybrid approach for high-fidelity digital humans. *Proceedings of the ACM on Computer Graphics and Interactive Techniques 6*, 3 (2023), 1–23. 19

[HML*14] HU L., MA C., LUO L., WEI L.-Y., LI H.: Capturing braided hairstyles. In *Transactions on Graphics, (Proc. SIGGRAPH Asia)* (2014), pp. 225:1–225:9. 13

[HMLL14] HU L., MA C., LUO L., LI H.: Robust hair capture using simulated examples. *Transactions on Graphics, (Proc. SIGGRAPH) 33*, 4 (2014). 13

[HMLL15] HU L., MA C., LUO L., LI H.: Single-view hair modeling using a hairstyle database. *Transactions on Graphics, (Proc. SIGGRAPH)* (2015), 125:1–125:9. 13, 32

[HPX*22] HONG Y., PENG B., XIAO H., LIU L., ZHANG J.: Headnerf: A real-time nerf-based parametric head model. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 20374–20384. 8, 10

[HSS*25] HE C., SUN X., SHU Z., LUAN F., PIRK S., HERRERA J. A. A., MICHELS D. L., WANG T. Y., ZHANG M., RUSHMEIER H., ZHOU Y.: Perm: A parametric representation for multi-style 3d hair modeling. In *International Conference on Learning Representations (ICLR)* (2025). 5, 13, 14

[HSW*17] 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. *Transactions on Graphics, (Proc. SIGGRAPH) 36*, 6 (2017). 13

[HSX*22] HALIMI O., STUYCK T., XIANG D., BAGAUTDINOV T. M., WEN H., KIMMEL R., SHIRATORI T., WU C., SHEIKH Y., PRADA F.: Pattern-based cloth registration and sparse-view animation. *Transactions on Graphics, (Proc. SIGGRAPH Asia) 41*, 6 (2022), 196–1. 16

[Hu24] HU L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 8153–8163. 16, 21

[HWP*23] HSU J., WANG T., PAN Z., GAO X., YUKSEL C., WU K.: Sag-free initialization for strand-based hybrid hair simulation. *Transactions on Graphics, (Proc. SIGGRAPH) 42*, 4 (2023). 29

[HWP*24] HSU J., WANG T., PAN Z., GAO X., YUKSEL C., WU K.: Real-time physically guided hair interpolation. *Transactions on Graphics, (Proc. SIGGRAPH) 43*, 4 (2024). 29

[HWS*25] HU L., WANG G., SHEN Z., GAO X., MENG D., ZHUO L., ZHANG P., ZHANG B., BO L.: Animate anyone 2: High-fidelity character image animation with environment affordance. In *International Conference on Computer Vision (ICCV)* (2025), pp. 1–11. 21

[HWW*18] HAN X., WU Z., WU Z., YU R., DAVIS L. S.: Viton: An image-based virtual try-on network. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 7543–7552. 31

[HXL*25] HONG F.-T., XU Z., LIU H., LIN Q., SONG L., SHU Z., ZHOU Y., CEYLAN D., XU D.: Free-viewpoint human animation with pose-correlated reference selection. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 26253–26262. 21

[HXSH23] HO H.-I., XUE L., SONG J., HILLIGES O.: Learning locally editable virtual humans. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 21024–21035. 18

[HYL*25] HUANG Y., YUAN Y., LI X., KAUTZ J., IQBAL U.: Adahuman: Animatable detailed 3d human generation with compositional multiview diffusion. In *International Conference on Computer Vision (ICCV)* (2025), pp. 1–11. 22

[HYW*23] HUANG L., YANG F., WEI C., CHEN Y. J. E., YUAN C., GAO M.: Towards realtime: A hybrid physics-based method for hair animation on gpu. *Computer Graphics and Interactive Techniques (PACM-CGIT) 6*, 3 (2023). 29

[HYZ*24] HE K., YAO K., ZHANG Q., YU J., LIU L., XU L.: Dress-code: Autoregressively sewing and generating garments from text guidance. *Transactions on Graphics, (Proc. SIGGRAPH) 43*, 4 (2024), 1–13. 17

[HZG*24] HONG Y., ZHANG K., GU J., BI S., ZHOU Y., LIU D., LIU F., SUNKAVALLI K., BUI T., TAN H.: Lrm: Large reconstruction model for single image to 3d. In *International Conference on Learning Representations (ICLR)* (2024). 10, 17

[HZS*24] HERRERA J. A. A., ZHOU Y., SUN X., SHU Z., HE C., PIRK S., MICHELS D. L.: Augmented mass-spring model for real-time dense hair simulation. In *International Conference on Computer Vision (ICCV)* (2024). 29

[HZZ*24] HU L., ZHANG H., ZHANG Y., ZHOU B., LIU B., ZHANG S., NIE L.: 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. 20

[ICN*23] IQBAL U., CALISKAN A., NAGANO K., KHAMIS S., MOLCHANOV P., KAUTZ J.: Rana: Relightable articulated neural avatars. In *International Conference on Computer Vision (ICCV)* (2023), pp. 23142–23153. 8, 19, 20

[IPS*25] IANDOLA F., PIDHORSKYI S., SANTESTEBAN I., GUPTA D., PAHUJA A., BARTOLOVIC N., YU F., GARBIN E., SIMON T., SAITO S.: Squeezeme: Mobile-ready distillation of gaussian full-body avatars. In *Transactions on Graphics, (Proc. SIGGRAPH)* (2025), pp. 1–11. 19

[IRG*23] IŞIK M., RÜNZ M., GEORGOPOULOS M., KHAKHULIN T., STARCK J., AGAPITO L., NIESSNER M.: Humanrf: High-fidelity neural radiance fields for humans in motion. *Transactions on Graphics, (Proc. SIGGRAPH) 42*, 4 (2023), 1–12. 32

[ISS*23] IWASE S., SAITO S., SIMON T., LOMBARDI S., BAGAUTDI-NOV T., JOSHI R., PRADA F., SHIRATORI T., SHEIKH Y., SARAGIH J.: Relightablehands: Efficient neural relighting of articulated hand models. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 16663–16673. 6, 8, 15, 16, 28

[JCSH23] JIANG T., CHEN X., SONG J., HILLIGES O.: Instantavatar: Learning avatars from monocular video in 60 seconds. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 16922–16932. 8, 20

[JGK*24] JIANG Z., GUO C., KAUFMANN M., JIANG T., VALENTIN J., HILLIGES O., SONG J.: Multiply: Reconstruction of multiple people from monocular video in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 109–118. 20

[JLL*24] JIN H., LI Y., LUAN F., XIANGLI Y., BI S., ZHANG K., XU Z., SUN J., SNAVELY N.: Neural gaffer: Relighting any object via diffusion. *Advances in Neural Information Processing Systems (NeurIPS) 37* (2024), 141129–141152. 28

[JPW*25] JIN Y., PENG S., WANG X., XIE T., XU Z., YANG Y., SHEN Y., BAO H., ZHOU X.: Diffuman4d: 4d consistent human view synthesis from sparse-view videos with spatio-temporal diffusion models. In *International Conference on Computer Vision (ICCV)* (2025), pp. 1–11. 19

[JSR*22] JAKOB W., SPEIERER S., ROUSSEL N., NIMIER-DAVID M., VICINI D., ZELTNER T., NICOLET B., CRESPO M., LEROY V., ZHANG Z.: Mitsuba 3 renderer, 2022. https://mitsuba-renderer.org. 28

[JSZ*25] JUNKAWITSCH H., SUN G., ZHU H., THEOBALT C., HABERMANN M.: Eva: Expressive virtual avatars from multi-view videos. In *SIGGRAPH Conference Papers (SA)* (2025), pp. 1–11. 19

[JYG*22] JI C., YU T., GUO K., LIU J., LIU Y.: Geometry-aware single-image full-body human relighting. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 388–405. 28

[JZH*20] JIANG B., ZHANG J., HONG Y., LUO J., LIU L., BAO H.: Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision (ECCV)* (2020), pp. 18–35. 12, 17, 32

[JZZ*25] JIANG J., ZENG W., ZHENG Z., YANG J., LIANG C., LIAO W., LIANG H., ZHANG Y., GAO M.: Omnihuman-1.5: Instilling an active mind in avatars via cognitive simulation. *arXiv* (2025). 21

[KAC*24] KOLOTOUROS N., ALLDIECK T., CORONA E., BAZAVAN E. G., SMINCHISESCU C.: Instant 3d human avatar generation using image diffusion models. In *European Conference on Computer Vision (ECCV)* (2024), Springer, pp. 177–195. 22

[Kaj86] KAJIYA J. T.: The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques* (1986), pp. 143–150. 27

[KAL*21] KARRAS T., AITTALA M., LAINE S., HÄRKÖNEN E., HELLSTEN J., LEHTINEN J., AILA T.: Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems (NeurIPS) 34* (2021), 852–863. 3, 15

[KAM24] KERBIRIOU G., AVRIL Q., MARCHAL M.: 3D Reconstruction and Semantic Modeling of Eyelashes. *Computer Graphics Forum (CGF)* (2024). 14

[KAZ*23] KOLOTOUROS N., ALLDIECK T., ZANFIR A., BAZAVAN E., FIERARU M., SMINCHISESCU C.: Dreamhuman: Animatable 3d avatars from text. *Advances in Neural Information Processing Systems (NeurIPS) 36* (2023), 10516–10529. 18

[KBM*24] KHIRODKAR R., BAGAUTDINOV T., MARTINEZ J., ZHAOEN S., JAMES A., SELEDNIK P., ANDERSON S., SAITO S.: Sapiens: Foundation for human vision models. In *European Conference on Computer Vision (ECCV)* (2024), Springer, pp. 206–228. 23

[KCF*22] KUANG Z., CHEN Y., FU H., ZHOU K., ZHENG Y.: DeepMVSHair: Deep hair modeling from sparse views. In *SIGGRAPH Asia Conference Papers (SA)* (2022). 12

[KE19] KANAMORI Y., ENDO Y.: Relighting humans: occlusion-aware inverse rendering for full-body human images. *Transactions on Graphics, (Proc. SIGGRAPH)* (2019). 28

[KGE*21] KAPPEL M., GOLYANIK V., ELGHARIB M., HENNINGSON J.-O., SEIDEL H.-P., CASTILLO S., THEOBALT C., MAGNOR M.: High-fidelity neural human motion transfer from monocular video.

In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 1541–1550. 20

[KGN24] KIRSCHSTEIN T., GIEBENHAIN S., NIESSNER M.: Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 5481–5492. 5, 6

[KGT*24] KIRSCHSTEIN T., GIEBENHAIN S., TANG J., GEORGOPOULOS M., NIESSNER M.: GGHead: Fast and Generalizable 3D Gaussian Heads. In *SIGGRAPH Asia Conference Papers (SA)* (2024), SA '24, Association for Computing Machinery. 10

[KHHB21] KOCABAS M., HUANG C.-H. P., HILLIGES O., BLACK M. J.: Pare: Part attention regressor for 3d human body estimation. In *International Conference on Computer Vision (ICCV)* (2021), pp. 11127–11137. 21

[KHTG24] KAIRANDA N., HABERMANN M., THEOBALT C., GOLYANIK V.: Neuralclothsim: Neural deformation fields meet the thin shell theory. *Advances in Neural Information Processing Systems (NeurIPS) 37* (2024), 110533–110572. 30

[KJJ*26] KI T., JANG S., JO J., YOON J., HWANG S. J.: Avatar forcing: Real-time interactive head avatar generation for natural conversation. *arXiv preprint arXiv:2601.00664* (2026). 23

[KK89] KAJIYA J. T., KAY T. L.: Rendering fur with three dimensional textures. *SIGGRAPH Conference Papers (SA)* (1989). 28

[KKLD23] KERBL B., KOPANAS G., LEIMKÜHLER T., DRETTAKIS G.: 3d gaussian splatting for real-time radiance field rendering. *Transactions on Graphics, (Proc. SIGGRAPH) 42*, 4 (2023), 139–1. 4, 7, 15, 19, 27

[KKSJ24] KIM T., KIM B., SAITO S., JOO H.: Gala: Generating animatable layered assets from a single scan. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 1535–1545. 18

[KL21] KOROSTELEVA M., LEE S.-H.: Generating datasets of 3d garments with sewing patterns. *arXiv* (2021). 32

[KL22] KOROSTELEVA M., LEE S.-H.: Neuraltailor: Reconstructing sewing pattern structures from 3d point clouds of garments. *Transactions on Graphics, (Proc. SIGGRAPH) 41*, 4 (2022), 1–16. 12, 17

[KLA19] KARRAS T., LAINE S., AILA T.: A style-based generator architecture for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 4401–4410. 10, 32

[KLF*23] KWON Y., LIU L., FUCHS H., HABERMANN M., THEOBALT C.: Deliffas: Deformable light fields for fast avatar synthesis. *Advances in Neural Information Processing Systems (NeurIPS) 36* (2023), 40944–40962. 19

[KLL*24] KARRAS J., LI Y., LIU N., ZHU L., YOO I., LUGMAYR A., LEE C., KEMELMACHER-SHLIZERMAN I.: Fashion-vdm: Video diffusion model for virtual try-on. In *SIGGRAPH Asia Conference Papers (SA)* (2024), pp. 1–11. 31

[KM17] KHUNGURN P., MARSCHNER S.: Azimuthal scattering from elliptical hair fibers. *Transactions on Graphics, (Proc. SIGGRAPH) 36*, 2 (2017). 28

[KMC24] KI T., MIN D., CHAE G.: Learning to generate conditional triplane for 3d-aware expression controllable portrait animation. In *European Conference on Computer Vision (ECCV)* (2024), Springer, pp. 476–493. 10

[KNS*25] KANSY M., NARUNIEC J., SCHROERS C., GROSS M., WEBER R. M.: Reenact anything: Semantic video motion transfer using motion-textual inversion. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers* (2025), pp. 1–12. 23

[KPBD19] KOLOTOUROS N., PAVLAKOS G., BLACK M. J., DANIILIDIS K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)* (2019), pp. 2252–2261. 21

[KPHT23] KARUNRATANAKUL K., PROKUDIN S., HILLIGES O., TANG S.: Harp: Personalized hand reconstruction from a monocular rgb video. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 12802–12813. 5, 6, 15, 16

[KPJD21] KOLOTOUROS N., PAVLAKOS G., JAYARAMAN D., DANILIDIS K.: Probabilistic modeling for human mesh recovery. In *International Conference on Computer Vision (ICCV)* (2021), pp. 11605–11614. 21

[KQG*23] KIRSCHSTEIN T., QIAN S., GIEBENHAIN S., WALTER T., NIESSNER M.: Nersemble: Multi-view radiance field reconstruction of human heads. *Transactions on Graphics, (Proc. SIGGRAPH) 42*, 4 (2023). 23, 32

[KRS*25] KIRSCHSTEIN T., ROMERO J., SEVASTOPOLSKY A., NIESSNER M., SAITO S.: Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. In *International Conference on Computer Vision (ICCV)* (2025). 8, 10

[KSLZ22] KHAKHULIN T., SKLYAROVA V., LEMPITSKY V., ZAKHAROV E.: Realistic one-shot mesh-based head avatars. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 345–362. 9

[KSN*25] KIM B., SAITO S., NAM G., SIMON T., SARAGIH J., JOO H., LI J.: Haircup: Hair compositional universal prior for 3d gaussian avatars. In *International Conference on Computer Vision (ICCV)* (2025). 8, 11, 13

[KTS*15] KHAMIS S., TAYLOR J., SHOTTON J., KESKIN C., IZADI S., FITZGIBBON A.: Learning an efficient model of hand shape variation from depth images. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 2540–2548. 15

[KTZ*24] KONG W., TIAN Q., ZHANG Z., MIN R., DAI Z., ZHOU J., XIONG J., LI X., WU B., ZHANG J., ET AL.: Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024). 5

[KWG*19] KULON D., WANG H., GÜLER R. A., BRONSTEIN M. M., ZAFEIRIOU S.: Single image 3d hand reconstruction with mesh convolutions. In *British Machine Vision Conference (BMVC)* (2019). 15

[KWK*25] KANT Y., WEBER E., KIM J. K., KHIRODKAR R., ZHAOEN S., MARTINEZ J., GILITSCHENSKI I., SAITO S., BAGAUTDINOV T.: Pippo: High-resolution multi-view humans from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 16418–16429. 21

[KZB*24] KABADAYI B., ZIELONKA W., BHATNAGAR B. L., PONS-MOLL G., THIES J.: Gan-avatar: Controllable personalized gan-based human head avatar. In *International Conference on 3D Vision (3DV)* (2024). 9

[Lam60] LAMBERT J. H.: *Photometria sive de mensura et gradibus luminis, colorum et umbrae*. sumptibus vidvae E. Klett, typis CP Detleffsen, 1760. 27

[LBB*17] 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 (2017), 194–1. 4, 5, 6, 7, 10, 11

[LBB*25] LABS B. F., BATIFOL S., BLATTMANN A., BOESEL F., CONSUL S., DIAGNE C., DOCKHORN T., ENGLISH J., ENGLISH Z., ESSER P., ET AL.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 11

[LCCZ22] LYU Q., CHAI M., CHEN X., ZHOU K.: Real-time hair simulation with neural interpolation. *Transactions on Visualization and Computer Graphics (TVCG) 28*, 4 (2022), 1894–1905. 29, 30

[LCD*25] LI R., CAO C., DUMERY C., YOU Y., LI H., FUA P.: Single view garment reconstruction using diffusion mapping via pattern coordinates. In *SIGGRAPH Conference Papers (SA)* (2025), pp. 1–11. 17

[LCL*24] LI Y., CHEN H.-Y., LARIONOV E., SARAFIANOS N., MATUSIK W., STUYCK T.: Diffavatar: Simulation-ready garment optimization with differentiable simulation. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 4368–4378. 17

[LCR24] LEROY V., CABON Y., REVAUD J.: Grounding image matching in 3d with mast3r, 2024. 4

[LCS*24a] LI J., CAO C., SCHWARTZ G., KHIRODKAR R., RICHARDT C., SIMON T., SHEIKH Y., SAITO S.: URAvatar: Universal relightable gaussian codec avatars. In *SIGGRAPH Asia Conference Papers (SA)* (2024), ACM, pp. 128:1–128:11. 8, 9, 28

[LCS*24b] LI J., CAO C., SCHWARTZ G., KHIRODKAR R., RICHARDT C., SIMON T., SHEIKH Y., SAITO S.: Uravatar: Universal relightable gaussian codec avatars. In *SIGGRAPH Asia Conference Papers (SA)* (2024), pp. 1–11. 28

[LDK*25] LU Y., DONG J., KWON Y., ZHAO Q., DAI B., DE LA TORRE F.: Gas: Generative avatar synthesis from a single image. In *International Conference on Computer Vision (ICCV)* (2025), pp. 1–11. 21

[LDN*25] LIU D., DENG T., NAM G., RONG Y., PIDHORSKYI S., LI J., SARAGIH J., METAXAS D. N., CAO C.: Lucas: Layered universal codec avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025). 11

[LDR14] LAGARDE S., DE ROUSIERS C.: Moving frostbite to physically based rendering 3.0. *SIGGRAPH Course: Physically Based Shading in Theory and Practice 3* (2014). 28

[LFL*23] LI X., FANG Y., LAN L., WANG H., YANG Y., LI M., JIANG C.: Subspace-preconditioned gpu projective dynamics with contact for cloth simulation. In *SIGGRAPH Asia Conference Papers (SA)* (2023), pp. 1–12. 30

[LHR*21] LIU L., HABERMANN M., RUDNEV V., SARKAR K., GU J., THEOBALT C.: Neural actor: Neural free-view synthesis of human actors with pose control. *Transactions on Graphics, (Proc. SIGGRAPH) 40*, 6 (2021), 1–16. 6, 15

[LHX*25] LIN Y., HONG Y., XU Z., LI X., XU C., SONG C., LI R., CHEN H., LAN J., ZHU H., ET AL.: Interanimate: Taming region-aware diffusion model for realistic human interaction animation. *arXiv (2025).* 21

[LJX*25] LI H., JIANG L., XIAO X., WANG T., YI H., WU B., CAI D.: Magicid: Hybrid preference optimization for id-consistent and dynamic-preserved video customization. In *ICCV* (2025), pp. 1–8. 21

[LJY*25] LIN G., JIANG J., YANG J., ZHENG Z., LIANG C.: Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *International Conference on Computer Vision (ICCV)* (2025), pp. 1–11. 5, 21

[LJZ*23] LI C., JIN L., ZHENG Y., YU Y., HAN X.: Ems: 3d eyebrow modeling from single-view images. *Transactions on Graphics, (Proc. SIGGRAPH Asia) 42*, 6 (2023). 14

[LKJ21] LI M., KAUFMAN D. M., JIANG C.: Codimensional incremental potential contact. *Transactions on Graphics, (Proc. SIGGRAPH) 40*, 4 (2021), 1–24. 17, 30

[LKK24] LEE J., KIM H., KWON T.: Learning-based self-collision avoidance in retargeting using body part-specific signed distance fields. 23

[LKL*25] LEE S., KIM S., LEE H., JEONG W.-S., LEE J. H.: Geoavatar: Geometrically-consistent multi-person avatar reconstruction from sparse multi-view videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 21138–21147. 19

[LL24] LI H., LIU X.: Strand-accurate multi-view facial hair reconstruction and tracking. *The Visual Computer (VC) 40*, 7 (2024), 4713–4724. 14

[LLJ*25] LI B., LI X., JIANG Y., XIE T., GAO F., WANG H., YANG Y., JIANG C.: Garmentdreamer: 3dgs guided garment synthesis with diverse geometry and texture details. In *International Conference on 3D Vision (3DV)* (2025). 18

[LLLL24] LIU H., LI C., LI Y., LEE Y. J.: Improved baselines with visual instruction tuning. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), IEEE, pp. 26286–26296. 14

[LLR13] LUO L., LI H., RUSINKIEWICZ S.: Structure-aware hair capture. *Transactions on Graphics, (Proc. SIGGRAPH) 32*, 4 (2013). 12, 13

[LLS*24] LIN S., LI Z., SU Z., ZHENG Z., ZHANG H., LIU Y.: Layga: Layered gaussian avatars for animatable clothing transfer. In *SIGGRAPH Conference Papers (SA)* (2024), pp. 1–11. 16, 19

[LLWL23] LIU H., LI C., WU Q., LEE Y. J.: Visual instruction tuning, 2023. 14

[LLZ25] LIN W., LI H., ZHU Y.: Controlhair: Physically-based video diffusion for controllable dynamic hair rendering, 2025. 14

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: SMPL: A skinned multi-person linear model. *Transactions on Graphics, (Proc. SIGGRAPH Asia) 34*, 6 (2015), 248:1–248:16. 4, 5, 6, 15, 16

[LOZ*24] LUO H., OUYANG M., ZHAO Z., JIANG S., ZHANG L., ZHANG Q., YANG W., XU L., YU J.: GaussianHair: Hair modeling and rendering with light-aware gaussians. *arXiv* (2024). 12, 13, 28, 29

[LQZ*25] LI H., QIU H., ZHANG S., WANG X., WEI Y., LI Z., ZHANG Y., WU B., CAI D.: Personalvideo: High id-fidelity video customization without dynamic and semantic degradation. In *International Conference on Computer Vision (ICCV)* (2025), pp. 19406–19416. 21

[LSS*19] LOMBARDI S., SIMON T., SARAGIH J., SCHWARTZ G., LEHRMANN A., SHEIKH Y.: Neural volumes: Learning dynamic renderable volumes from images. *Transactions on Graphics, (Proc. SIGGRAPH) 38*, 4 (2019), 65:1–65:14. 7

[LSS*21] LOMBARDI S., SIMON T., SCHWARTZ G., ZOLLHOEFER M., SHEIKH Y., SARAGIH J.: Mixture of volumetric primitives for efficient neural rendering. *Transactions on Graphics, (Proc. SIGGRAPH) 40*, 4 (2021). 6, 7, 28

[LSS*23] LI J., SAITO S., SIMON T., LOMBARDI S., LI H., SARAGIH J.: Megane: Morphable eyeglass and avatar network. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 12769–12779. 28

[LSSS18] LOMBARDI S., SARAGIH J., SIMON T., SHEIKH Y.: Deep appearance models for face rendering. *Transactions on Graphics, (Proc. SIGGRAPH) 37*, 4 (2018), 1–13. 7

[LSZ*23] LI Z., SUN Y., ZHENG Z., WANG L., ZHANG S., LIU Y.: Animatable and relightable gaussians for high-fidelity human avatar modeling. *arXiv* (2023). 19

[LSZ*25] LI X., SUN Q., ZHANG P., YE F., LIAO Z., FENG W., ZHAO S., HE Q.: Anydressing: Customizable multi-garment virtual dressing via latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), IEEE, pp. 23723–23733. 31

[LVKS21] LEWIS K. M., VARADHARAJAN S., KEMELMACHER-SHLIZERMAN I.: Tryongan: Body-aware try-on via layered interpolation. *Transactions on Graphics (TOG) 40*, 4 (2021), 1–10. 31

[LWH*25] LIU J., WANG J., HOU S., REN M., WU H., MA L., PEI R., HE Z.: Beyond face swapping: A diffusion-based digital human benchmark for multimodal deepfake detection. *arXiv* (2025). 22

[LXH*20] LIU L., XU W., HABERMANN M., ZOLLHÖFER M., BERNARD F., KIM H., WANG W., THEOBALT C.: Neural human video rendering by learning dynamic textures and rendering-to-video translation. *TVCG* (2020), 1–10. 20

[LXL*23] LIU L., XU X., LIN Z., LIANG J., YAN S.: Towards garment sewing pattern reconstruction from a single image. *Transactions on Graphics, (Proc. SIGGRAPH Asia) 42*, 6 (2023), 1–15. 17, 18

[LXL*25a] LI Y., XIA M., LIU G., BAI J., WANG X., ZHANG C., LIN Y., CHU R., WAN P., YANG Y.: Adaviewplanner: Adapting video diffusion models for viewpoint planning in 4d scenes. *arXiv preprint arXiv:2510.10670* (2025). 5, 23

[LXL*25b] LIAO Z., XU Y., LI Z., LI Q., ZHOU B., BAI R., XU D., ZHANG H., LIU Y.: Hhavatar: Gaussian head avatar with dynamic hairs. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI) PP* (2025). 29, 30

[LXS*25] LI R., XING D., SUN H., HA Y., SHEN J., HO C.: Tokenmotion: Decoupled motion control via token disentanglement for human-centric video generation. In *CVPR* (2025), pp. 1951–1961. 5, 21, 23

[LXZ*19] LIU L., XU W., ZOLLHOEFER M., KIM H., BERNARD F., HABERMANN M., WANG W., THEOBALT C.: Neural rendering and reenactment of human actor videos. *TOG* (2019), 1–14. 20

[LYD*25] LI X., YU C., DU W., JIANG Y., XIE T., CHEN Y., YANG Y., JIANG C.: Dress-1-to-3: Single image to simulation-ready 3d outfit with diffusion prior and differentiable physics. *Transactions on Graphics, (Proc. SIGGRAPH) 44*, 4 (2025), 1–16. 18

[LYDM*25] LI X., YUAN Y., DE MELLO S., DAVIET G., LEAF J., MACKLIN M., KAUTZ J., IQBAL U.: Simavatar: Simulation-ready avatars with layered hair and clothing. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 26320–26330. 12, 18

[LZL*23] LI Z., ZHENG Z., LIU Y., ZHOU B., LIU Y.: Posevocab: Learning joint-structured pose embeddings for human avatar modeling. In *SIGGRAPH Conference Papers (SA)* (2023), pp. 1–11. 19

[LZL*25a] LI P., ZHENG W., LIU Y., YU T., LI Y., QI X., CHI X., XIA S., CAO Y.-P., XUE W., ET AL.: Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 16008–16018. 22

[LZL*25b] LIANG J., ZHOU J., LI S., CAO C., SUN L., QIAN Y., CHEN W., WANG F.: Realismotion: Decomposed human motion control and video generation in the world space. *arXiv preprint arXiv:2508.08588* (2025). 5, 23

[LZQ*22] LI Y., ZHANG L., QIU Z., JIANG Y., LI N., MA Y., ZHANG Y., XU L., YU J.: Nimble: a non-rigid hand model with bones and muscles. *Transactions on Graphics, (Proc. SIGGRAPH) 41*, 4 (2022), 1–16. 1, 8, 15, 16

[LZWL24] LI Z., ZHENG Z., WANG L., LIU Y.: Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 19711–19722. 8, 19

[LZY*25] LI D., ZHONG W., YU W., PAN Y., ZHANG D., YAO T., HAN J., MEI T.: Pursuing temporal-consistent video virtual try-on via dynamic pose interaction. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 22648–22657. 31

[LZZ*22] LIN S., ZHANG H., ZHENG Z., SHAO R., LIU Y.: Learning implicit templates for point-based clothed human modeling. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 210–228. 18

[may] Autodesk, INC. maya. Maya. Accessed: 2025-10-10. 11, 28, 29

[MBT*12] MIGUEL E., BRADLEY D., THOMASZEWSKI B., BICKEL B., MATUSIK W., OTADUY M. A., MARSCHNER S.: Data-driven estimation of cloth simulation models. *Computer Graphics Forum (CGF) 31*, 2pt2 (2012), 519–528. 17

[MBZ*22] MIHAJLOVIC M., BANSAL A., ZOLLHOEFER M., TANG S., SAITO S.: Keypointnerf: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 179–197. 20

[MCH*24] MEISHVILI G., CLEMOES J., HEWITT C., HOSENIE Z., XIAO X., DE LA GORCE M., TAKACS T., BALTRUSAITIS T., CRIMINISI A., MCRAE C., JABLONSKI N., WILCZKOWIAK M.: Hairmony: Fairness-aware hairstyle classification. In *SIGGRAPH Asia Conference Papers (SA)* (2024). 13

[MDT*25] MENDIRATTA M., DESHMUKH M., TEOTIA K., GOLYANIK V., KORTYLEWSKI A., THEOBALT C.: Grmm: Real-time high-fidelity gaussian morphable head model with learned residuals. *arXiv* (2025). 9, 10

[MFH*25] MA Y., FENG K., HU Z., WANG X., WANG Y., ZHENG M., HE X., ZHU C., LIU H., HE Y., WANG Z., LI Z., LI X., LIU W., XU D., ZHANG L., CHEN Q.: Controllable video generation: A survey, 2025. 2, 16

[MHHR07] MÜLLER M., HEIDELBERGER B., HENNIX M., RATCLIFF J.: Position based dynamics. *Journal of Visual Communication and Image Representation 18*, 2 (2007), 109–118. 30

[MHM*25] MEI Y., HE M., MA L., PHILIP J., XIAN W., GEORGE D. M., YU X., DEDIC G., TAŞEL A. L., YU N., PATEL V. M., DEBEVEC P.: Lux post facto: Learning portrait performance relighting with conditional video diffusion and a hybrid dataset. *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025). 9

[MJC*03] MARSCHNER S. R., JENSEN H. W., CAMMARANO M., WORLEY S., HANRAHAN P.: Light scattering from human hair fibers. *Transactions on Graphics, (Proc. SIGGRAPH) 22*, 3 (2003), 780–791. 6, 28

[MKR*24] MARTINEZ J., KIM E., ROMERO J., BAGAUTDINOV T., SAITO S., YU S.-I., ANDERSON S., ZOLLHÖFER M., WANG T.-L., BAI S., ET AL.: Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. *Advances in Neural Information Processing Systems (NeurIPS) 37* (2024), 83008–83023. 32

[MLW*24] MA Y., LIU H., WANG H., PAN H., HE Y., YUAN J., ZENG A., CAI C., SHUM H.-Y., LIU W., ET AL.: Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. *arXiv* (2024). 5, 11

[MMC16] MACKLIN M., MÜLLER M., CHENTANEZ N.: Xpbd: Position-based simulation of compliant constrained dynamics. In *Proceedings of the 9th International Conference on Motion in Games (MiG)* (2016), pp. 49–54. 30

[MNSL22] MOON G., NAM H., SHIRATORI T., LEE K. M.: 3d clothed human reconstruction in the wild. In *European Conference on Computer Vision (ECCV)* (2022), pp. 184–200. 17

[MPS21] MORALES A., PIELLA G., SUKNO F. M.: Survey on 3d face reconstruction from uncalibrated images. *Computer Science Review 40* (2021), 100400. 2

[MSD*24] MOREAU A., SONG J., DHAMO H., SHAW R., ZHOU Y., PÉREZ-PELLITERO E.: Human gaussian splatting: Real-time rendering of animatable avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 788–798. 8, 19

[MSL20] MOON G., SHIRATORI T., LEE K. M.: Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *European Conference on Computer Vision (ECCV)* (2020), Springer, pp. 440–455. 15

[MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)* (2020), Springer. 4, 18, 20

[MSY*21] MA Q., SAITO S., YANG J., TANG S., BLACK M. J.: Scale: Modeling clothed humans with a surface codec of articulated local elements. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 16082–16093. 18

[Mül06] MÜLLER C.: *Spherical harmonics*, vol. 17. Springer, 2006. 27

[MWH*23] MUNDRA A., WANG J., HABERMANN M., THEOBALT C., ELGHARIB M., ET AL.: Livehand: Real-time and photorealistic neural hand rendering. In *International Conference on Computer Vision (ICCV)* (2023), pp. 18035–18045. 15

[MXJ*24] MOON G., XU W., JOSHI R., WU C., SHIRATORI T.: Authentic hand avatar from a phone scan via universal hand model. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 2029–2038. 15

[MYCB25] MEN Y., YAO Y., CUI M., BO L.: Mimo: Controllable character video synthesis with spatial decomposed modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 21181–21191. 21

[MYR*20] MA Q., YANG J., RANJAN A., PUJADES S., PONS-MOLL G., TANG S., BLACK M. J.: Learning to dress 3d people in generative clothing. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 6469–6478. 18

[MYTB21] MA Q., YANG J., TANG S., BLACK M. J.: The power of points for modeling humans in clothing. In *International Conference on Computer Vision (ICCV)* (2021), pp. 10974–10984. 18

[MYW*20] MOON G., YU S.-I., WEN H., SHIRATORI T., LEE K. M.: InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *European Conference on Computer Vision (ECCV)* (2020). 15, 24, 32

[MZBT21] MIHAJLOVIC M., ZHANG Y., BLACK M. J., TANG S.: Leap: Learning articulated occupancy of people. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 10461–10471. 18

[MZQ*23] MA Z., ZHU X., QI G.-J., LEI Z., ZHANG L.: Otavatar: One-shot talking face avatar with controllable tri-plane rendering. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 16901–16910. 5, 10

[NAK*25] NAKAYAMA K., ACKERMANN J., KESDOGAN T. L., ZHENG Y., KOROSTELEVA M., SORKINE-HORNUNG O., GUIBAS L. J., YANG G., WETZSTEIN G.: Aipparel: A multimodal foundation model for digital garments. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 8138–8149. 17

[NBC*24] NARASIMHASWAMY S., BHATTACHARYA U., CHEN X., DASGUPTA I., MITRA S., HOAI M.: Handiffuser: Text-to-image generation with realistic hand appearances. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 2468–2479. 16

[NCZ17] NAGRANI A., CHUNG J. S., ZISSERMAN A.: Voxceleb: a large-scale speaker identification dataset. *CoRR* (2017). 32

[NHZ*22] NEEKHARA P., HUSSAIN S., ZHANG X., HUANG K., MCAULEY J., KOUSHANFAR F.: Facesigns: semi-fragile neural watermarks for media authentication and countering deepfakes. *arXiv* (2022). 22

[NKML25] NAM H., KIM D., MOON G., LEE K. M.: Parte: Part-guided texturing for 3d human reconstruction from a single image. In *International Conference on Computer Vision (ICCV)* (2025), pp. 1–11. 21

[NKOL25] NAM H., KIM D., OH J., LEE K. M.: Decloth: Decomposable 3d cloth and human body reconstruction from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 5636–5645. 17

[NWKS19] NAM G., WU C., KIM M. H., SHEIKH Y.: Strand-accurate multi-view hair capture. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 155–164. 12

[OBT25] OSTREK M., BLACK M. J., THIES J.: Hairfree: Compositional 2d head prior for text-driven 360° bald texture synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)* (2025). 11

[ODM*23] OQUAB M., DARCET T., MOUTAKANNI T., VO H. Q., SZAFRANIEC M., KHALIDOV V., FERNANDEZ P., HAZIZA D., MASSA F., EL-NOUBY A., ASSRAN M., BALLAS N., GALUBA W., HOWES R., HUANG P.-Y. B., LI S.-W., MISRA I., RABBAT M. G., SHARMA V., SYNNAEVE G., XU H., JÉGOU H., MAIRAL J., LABATUT P., JOULIN A., BOJANOWSKI P.: Dinov2: Learning robust visual features without supervision. *ArXiv* (2023). 13

[OKA11a] OIKONOMIDIS I., KYRIAZIS N., ARGYROS A.: Efficient model-based 3d tracking of hand articulations using kinect. In *British Machine Vision Conference (BMVC)* (2011). 14

[OKA11b] OIKONOMIDIS I., KYRIAZIS N., ARGYROS A. A.: Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *International Conference on Computer Vision (ICCV)* (2011), IEEE, pp. 2088–2095. 14

[OLS*25] OHKAWA T., LEE J., SAITO S., SARAGIH J., PRADA F., XU Y., YU S.-I., FURUTA R., SATO Y., SHIRATORI T.: Generative modeling of shape-dependent self-contact human poses. In *International Conference on Computer Vision (ICCV)* (2025), pp. 5426–5436. 16

[oSN77] OF STANDARDS U. S. N. B., NICODEMUS F. E.: *Geometrical considerations and nomenclature for reflectance*, vol. 160. US Department of Commerce, National Bureau of Standards Washington, DC, USA, 1977. 27

[OT24]   OSTREK M., THIES J.: Stable video portraits. In *European Conference on Computer Vision (ECCV)* (2024). 2

[PBD*10]   PARKER S. G., BIGLER J., DIETRICH A., FRIEDRICH H., HOBEROCK J., LUEBKE D., MCALLISTER D., MCGUIRE M., MORLEY K., ROBISON A., STICH M.: Optix: a general purpose ray tracing engine. *Transactions on Graphics (TOG) 29*, 4 (2010). 28

[PBV23]   PETROVICH M., BLACK M. J., VAROL G.: TMR: text-to-motion retrieval using contrastive 3D human motion synthesis. In *International Conference on Computer Vision (ICCV)* (2023), pp. 9488–9497. 6

[PCG*19]   PAVLAKOS G., CHOUTAS V., GHORBANI N., BOLKART T., OSMAN A. A., TZIONAS D., BLACK M. J.: Expressive body capture: 3d hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 10975–10985. 5, 6, 16, 18, 20

[PDW*21]   PENG S., DONG J., WANG Q., ZHANG S., SHUAI Q., ZHOU X., BAO H.: Animatable neural radiance fields for modeling dynamic human bodies. In *International Conference on Computer Vision (ICCV)* (2021), pp. 14314–14323. 18

[PFS*19]   PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: Deepsdf: Learning continuous signed distance functions for shape representation. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 165–174. 10

[PHBB*24]   PARKER-HOLDER J., BALL P., BRUCE J., DASAGI V., HOLSHEIMER K., KAPLANIS C., MOUFAREK A., SCULLY G., SHAR J., SHI J., ET AL.:   Genie 2: A large-scale foundation world model.   *URL: https://deepmind. google/discover/blog/genie-2-a-large-scale-foundation-world-model* (2024). 23

[PJBM23]   POOLE B., JAIN A., BARRON J. T., MILDENHALL B.: Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)* (2023). 5, 17, 18

[PJH23]   PHARR M., JAKOB W., HUMPHREYS G.: *Physically based rendering: From theory to implementation.* MIT Press, 2023. 27

[PKA*09]   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 (AAAI)* (2009), Ieee, pp. 296–301. 5, 10

[PKP*24]   PAUDEL P., KHANAL A., PAUDEL D. P., TANDUKAR J., CHHATKULI A.:   ihuman: Instant animatable digital humans from monocular videos. In *European Conference on Computer Vision (ECCV)* (2024), Springer, pp. 304–323. 20

[PLPM20]   PATEL C., LIAO Z., PONS-MOLL G.: Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 7365–7375. 30

[PLT*25]   PANG Y., LIU J., TAN L., ZHANG Y., GAO F., DENG X., KANG Z., WEI X., LIU Y.: Mavid: A multimodal framework for audio-visual dialogue understanding and generation. *arXiv* (2025). 21

[PMJ*22]   PAN X., MAI J., JIANG X., TANG D., LI J., SHAO T., ZHOU K., JIN X., MANOCHA D.: Predicting loose-fitting garment deformations using bone-driven motion networks. In *SIGGRAPH Conference Papers (SA)* (2022), pp. 1–10. 17, 30

[PMPHB17]   PONS-MOLL G., PUJADES S., HU S., BLACK M. J.: Clothcap: Seamless 4d clothing capture and retargeting. *Transactions on Graphics, (Proc. SIGGRAPH) 36*, 4 (2017), 1–15. 16

[PNK25]   PAN Y., NIESSNER M., KIRSCHSTEIN T.: Hairgs: Hair strand reconstruction based on 3d gaussian splatting. *arXiv* (2025). 13

[PPM*23]   POTAMIAS R. A., PLOUMPIS S., MOSCHOGLOU S., TRIANTAFYLLOU V., ZAFEIRIOU S.: Handy: Towards a high fidelity 3d hand shape and appearance model. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 4670–4680. 5, 15

[PSB*21]   PARK K., SINHA U., BARRON J. T., BOUAZIZ S., GOLDMAN D. B., SEITZ S. M., MARTIN-BRUALLA R.: Nerfies: Deformable neural radiance fields. In *International Conference on Computer Vision (ICCV)* (2021), pp. 5865–5874. 4, 17

[PSH*21]   PARK K., SINHA U., HEDMAN P., BARRON J. T., BOUAZIZ S., GOLDMAN D. B., MARTIN-BRUALLA R., SEITZ S. M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *Transactions on Graphics, (Proc. SIGGRAPH) 40* (2021). 4

[PSZ*25]   PANG Y., SHAO R., ZHANG J., TU H., LIU Y., ZHOU B., ZHANG H., LIU Y.: Manivideo: Generating hand-object manipulation video with dexterous and generalizable grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 12209–12219. 21

[PWS*23]   PENG Z., WU H., SONG Z., XU H., ZHU X., HE J., LIU H., FAN Z.: EmoTalk: Speech-driven emotional disentanglement for 3D face animation. In *International Conference on Computer Vision (ICCV)* (2023), IEEE, pp. 20630–20640. 6

[PX23]   PEEBLES W., XIE S.: Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (2023), pp. 4195–4205. 31

[PYG*24]   PO R., YIFAN W., GOLYANIK V., ABERMAN K., BARRON J. T., BERMANO A., CHAN E., DEKEL T., HOLYNSKI A., KANAZAWA A., ET AL.:   State of the art on diffusion models for visual computing. *Computer Graphics Forum (CGF) 43*, 2 (2024), e15063. 2

[PZK*24]   PANG H., ZHU H., KORTYLEWSKI A., THEOBALT C., HABERMANN M.:   Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 1165–1175. 19

[PZS*25]   PRINZLER M., ZAKHAROV E., SKLYAROVA V., KABADAYI B., THIES J.: Joker: Conditional 3d head synthesis with extreme facial expressions. In *International Conference on 3D Vision (3DV)* (2025), IEEE, pp. 1583–1593. 11

[PZX*21]   PENG S., ZHANG Y., XU Y., WANG Q., SHUAI Q., BAO H., ZHOU X.: 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)* (2021), pp. 9054–9063. 6, 8, 18, 24, 32

[QCZ*23]   QIU L., CHEN G., ZHOU J., XU M., WANG J., HAN X.: Rec-mv: Reconstructing 3d dynamic cloth from monocular videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 4637–4646. 12, 17

[QGL*25]   QIU L., GU X., LI P., ZUO Q., SHEN W., ZHANG J., QIU K., YUAN W., CHEN G., DONG Z., ET AL.: Lhm: Large animatable human reconstruction model from a single image in seconds. *arXiv* (2025). 23

[QKS*24]   QIAN S., KIRSCHSTEIN T., SCHONEVELD L., DAVOLI D., GIEBENHAIN S., NIESSNER M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 20299–20309. 1, 5, 6, 7, 8, 9

[QLZ*25]   QIU L., LI P., ZUO Q., GU X., DONG Y., YUAN W., HAN X., CHEN G., DONG Z.: Pf-lhm: 3d animatable avatar reconstruction from pose-free articulated human images. In *International Conference on Computer Vision (ICCV)* (2025), pp. 1–11. 21

[QWM*20]   QIAN N., WANG J., MUELLER F., BERNARD F., GOLYANIK V., THEOBALT C.: Html: A parametric hand texture model for 3d hand reconstruction and personalization. In *European Conference on Computer Vision (ECCV)* (2020), Springer, pp. 54–71. 15

[QXL*22]   QIAN S., XU J., LIU Z., MA L., GAO S.: Unif: United neural implicit functions for clothed human reconstruction and animation. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 121–137. 18

[QZZ*25]   QIU L., ZHU S., ZUO Q., GU X., DONG Y., ZHANG J., XU C., LI Z., YUAN W., BO L., ET AL.: Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 21148–21158. 8, 19, 22, 23

[RAFA*25] RESSLER-ANTAL T., FUNDEL F., ALAYA M. B., BAU-MANN S. A., KRAUSE F., GUI M., OMMER B.: Dismo: Disentangled motion representations for open-world motion transfer. *arXiv preprint arXiv:2511.23428* (2025). 23

[Ram02] RAMAMOORTHI R.: Analytic pca construction for theoretical analysis of lighting variability in images of a lambertian object. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 24*, 10 (2002), 1322–1333. 28

[RBS*22] REMELLI E., BAGAUTDINOV T., SAITO S., WU C., SIMON T., WEI S.-E., GUO K., CAO Z., PRADA F., SARAGIH J., ET AL.: Drivable volumetric avatars using texel-aligned features. In *SIGGRAPH Conference Papers (SA)* (2022), pp. 1–9. 6, 8, 18

[RCV*19] ROSSLER A., COZZOLINO D., VERDOLIVA L., RIESS C., THIES J., NIESSNER M.: Faceforensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)* (2019), pp. 1–11. 22

[RGB*20] RIVIERE J., GOTARDO P. F., BRADLEY D., GHOSH A., BEELER T.: Single-shot high-quality facial geometry and skin appearance capture. *Transactions on Graphics, (Proc. SIGGRAPH) 39*, 4 (2020), 81. 28

[RGW*25] RONG B., GRIGOREV A., WANG W., BLACK M. J., THOMASZEWSKI B., TSALICOGLOU C., HILLIGES O.: Gaussian garments: Reconstructing simulation-ready clothing with photorealistic appearance from multi-view video. *International Conference on 3D Vision (3DV)* (2025). 12, 16, 17

[RH01] RAMAMOORTHI R., HANRAHAN P.: An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* (2001), pp. 497–500. 6, 27

[RK94] REHG J. M., KANADE T.: Visual tracking of high dof articulated structures: an application to human hand tracking. In *European Conference on Computer Vision (ECCV)* (1994), Springer, pp. 35–46. 14

[RMBCO22] ROICH D., MOKADY R., BERMANO A. H., COHEN-OR D.: Pivotal tuning for latent-based editing of real images. *Transactions on Graphics, (Proc. SIGGRAPH) 42*, 1 (2022), 1–13. 10

[RMZ*25] RAO P., MEKA A., ZHOU X., FOX G., B R M., ZHAN F., WEYRICH T., BICKEL B., PFISTER H., MATUSIK W., BEELER T., EL-GHARIB M., HABERMANN M., THEOBALT C.: 3DPR: Single image 3D portrait relighting with generative priors. In *SIGGRAPH Asia Conference Papers (SA)* (2025), pp. 1–11. 9

[RSW*22] ROSU R. A., SAITO S., WANG Z., WU C., BEHNKE S., NAM G.: Neural strands: Learning hair geometry and appearance from multi-view images. In *European Conference on Computer Vision (ECCV)* (2022). 12, 28, 29

[RTB17] ROMERO J., TZIONAS D., BLACK M. J.: Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics, (Proc. SIGGRAPH Asia) 36*, 6 (Nov. 2017). 5, 6, 15

[RWF*25] ROSU R. A., WU K., FENG Y., ZHENG Y., BLACK M. J.: DiffLocks: Generating 3d hair from a single image using diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025). 1, 12, 13, 14

[RYCT23] RANJAN A., YI K. M., CHANG J.-H. R., TUZEL O.: Facelit: Neural 3d relightable faces. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 8619–8628. 28

[RYV*25] R. M. B., YIN F., VOLETI V., DROBYSHEV N., LAPIN M., VASISHTA A., JAMPANI V.: Stable video-driven portraits, 2025. 5, 11

[SAH*25] SALEH F., ALIAKBARIAN S., HEWITT C., PETIKAM L., XIAO-XIAN, CRIMINISI A., CASHMAN T. J., BALTRUŠAITIS T.: DAViD: Data-efficient and accurate vision models from synthetic data. In *International Conference on Computer Vision (ICCV)* (2025). 23

[SBR23a] SU S.-Y., BAGAUTDINOV T., RHODIN H.: Npc: Neural point characters from video. In *International Conference on Computer Vision (ICCV)* (2023), pp. 14795–14805. 20

[SBR23b] SU S.-Y., BAGAUTDINOV T., RHODIN H.: Npc: Neural point characters from video. In *International Conference on Computer Vision (ICCV)* (2023), pp. 1–8. 20

[SC23] STUYCK T., CHEN H.-Y.: Diffxpbd: Differentiable position-based simulation of compliant constraint dynamics. *Symposium on Computer Animation (SCA) 6*, 3 (2023), 1–14. 17

[SCD*23] SKLYAROVA V., CHELISHEV J., DOGARU A., MEDVEDEV I., LEMPITSKY V., ZAKHAROV E.: Neural haircut: Prior-guided strand-based hair reconstruction. In *International Conference on Computer Vision (ICCV)* (2023), pp. 19762–19773. 12, 13, 28, 29

[SCT*25] SUN X., CAI Z., TAI Y., YANG J., ZHANG Z.: Strandhead: Text to hair-disentangled 3d head avatars using human-centric priors. In *International Conference on Computer Vision (ICCV)* (2025). 14

[SDF*24] SUN G., DABRAL R., FUA P., THEOBALT C., HABERMANN M.: MetaCap: Meta-learning priors from multi-view imagery for sparse-view human performance capture and rendering. In *European Conference on Computer Vision (ECCV)* (2024), pp. 341–361. 19

[SDZ*25] SUN G., DABRAL R., ZHU H., FUA P., THEOBALT C., HABERMANN M.: Real-time free-view human rendering from sparse-view RGB videos using double unprojected textures. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 562–573. 19

[SF16] SCHÖNBERGER J. L., FRAHM J.-M.: Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2016). 4, 13

[SGBL23] SVITOV D., GUDKOV D., BASHIROV R., LEMPITSKY V.: Dinar: Diffusion inpainting of neural textures for one-shot human avatars. In *International Conference on Computer Vision (ICCV)* (2023), pp. 7062–7072. 21

[SGK*23] SHEN K., GUO C., KAUFMANN M., ZARATE J. J., VALENTIN J., SONG J., HILLIGES O.: X-avatar: Expressive human avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 16911–16921. 18

[SGN25] SCHMIDT J., GIEBENHAIN S., NIESSNER M.: Becominglit: Relightable gaussian avatars with hybrid neural shading, 2025. 9, 27, 28

[SGPT23] SHIMADA S., GOLYANIK V., PÉREZ P., THEOBALT C.: Decaf: Monocular deformation capture for face and hand interactions. *Transactions on Graphics (TOG) 42*, 6 (2023), 1–16. 16

[SGY*24] SANYAL S., GHOSH P., YANG J., BLACK M. J., THIES J., BOLKART T.: SCULPT: Shape-conditioned unpaired learning of pose-dependent clothed and textured human meshes. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 2362–2371. 21

[SHL*23] SU Z., HU L., LIN S., ZHANG H., ZHANG S., THIES J., LIU Y.: Caphy: Capturing physical properties for animatable human avatars. In *International Conference on Computer Vision (ICCV)* (2023), pp. 14150–14160. 17, 18

[SHM*18] SAITO S., HU L., MA C., IBAYASHI H., LUO L., LI H.: 3d hair synthesis using volumetric variational autoencoders. *Transactions on Graphics, (Proc. SIGGRAPH) 37*, 6 (2018). 12, 13

[SHM*25] SHAN M., HE Z., MA H., JUEFEI-XU F., ZHANG P., HOU T., CHUANG C.-Y.: Populate-a-scene: Affordance-aware human video generation. *arXiv* (2025). 21

[SHN*19] SAITO S., HUANG Z., NATSUME R., MORISHIMA S., KANAZAWA A., LI H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision (ICCV)* (2019), pp. 2304–2314. 21

[SHS*24] SHETTY A., HABERMANN M., SUN G., LUVIZON D., GOLYANIK V., THEOBALT C.: Holoported characters: Real-time free-viewpoint rendering of humans from sparse RGB cameras. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 1206–1215. 19

[SKS02] Sloan P. J., Kautz J., Snyder J. M.: Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. *Transactions on Graphics (TOG) 21*, 3 (2002), 527–536. 27

[SLB*22] Sun Y., Liu W., Bao Q., Fu Y., Mei T., Black M. J.: Putting people in their place: Monocular regression of 3d people in depth. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 13243–13252. 21

[SLD23] Shao Y., Loy C. C., Dai B.: Towards multi-layered 3d garments animation. In *International Conference on Computer Vision (ICCV)* (2023), pp. 14361–14370. 30

[SLD24] Shao Y., Loy C. C., Dai B.: Learning 3d garment animation from trajectories of a piece of cloth. In *Advances in Neural Information Processing Systems (NeurIPS)* (2024), pp. 41803–41825. 30

[SLH*19] Smith D., Loper M., Hu X., Mavroidis P., Romero J.: Facsimile: Fast and accurate scans from an image in less than a second. In *International Conference on Computer Vision (ICCV)* (2019), pp. 5330–5339. 21

[SLL*24a] Shin J., Lee J., Lee S., Park M.-G., Kang J.-M., Yoon J. H., Jeon H.-G.: Canonicalfusion: Generating drivable 3d human avatars from multiple images. In *European Conference on Computer Vision (ECCV)* (2024), Springer, pp. 38–56. 20

[SLL*24b] Stuyck T., Lin G. W.-C., Larionov E., Chen H.-Y., Bozic A., Sarafianos N., Roble D.: Quaffure: Real-time quasi-static neural hair simulation. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 29, 30

[SM25] Sim G., Moon G.: Persona: Personalized whole-body 3d avatar with pose-driven deformations from a single image. In *International Conference on Computer Vision (ICCV)* (2025), pp. 1–11. 8, 22

[SMOT15] Sridhar S., Mueller F., Oulasvirta A., Theobalt C.: Fast and robust hand tracking using detection-guided optimization. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3213–3221. 14

[SNA*21] Sun T., Nam G., Aliaga C., Hery C., Ramamoorthi R.: Human Hair Inverse Rendering using Multi-View Photometric data. In *Computer Graphics Forum (CGF)* (2021). 28, 29

[SNF14] Schmidt T., Newcombe R. A., Fox D.: Dart: Dense articulated real-time tracking. *Robotics: Science and Systems (RSS) 2*, 1 (2014), 1–9. 14

[SOC19] Santesteban I., Otaduy M. A., Casas D.: Learning-based animation of clothing for virtual try-on. *Computer Graphics Forum (CGF) 38*, 2 (2019), 355–366. 30

[SOC22] Santesteban I., Otaduy M. A., Casas D.: Snug: Self-supervised neural dynamic garments. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 8140–8150. 17, 30

[SOT13] Sridhar S., Oulasvirta A., Theobalt C.: Interactive markerless articulated hand motion tracking using rgb and depth data. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2013), pp. 2456–2463. 14

[SPJT10] Sadeghi I., Pritchett H., Jensen H. W., Tamstorf R.: An artist friendly hair shading system. *Transactions on Graphics, (Proc. SIGGRAPH) 29*, 4 (2010). 28

[SPZ*24] Shao R., Pang Y., Zheng Z., Sun J., Liu Y.: 360-degree human video generation with 4d diffusion transformer. *Transactions on Graphics, (Proc. SIGGRAPH) 43*, 6 (2024), 1–13. 32

[SRRW21] Sztrajman A., Rainer G., Ritschel T., Weyrich T.: Neural brdf representation and importance sampling. *Computer Graphics Forum (CGF) 40*, 6 (2021), 332–346. 27

[SS99] Stockman A., Sharpe L. T.: Cone spectral sensitivities and color matching. *Color vision: From genes to perception* (1999), 53–88. 27

[SSK*05] Scholz V., Stich T., Keckeisen M., Wacker M., Magnor M.: Garment motion capture using color-coded patterns. *Computer Graphics Forum (CGF) 24*, 3 (2005), 439–448. 16

[SSS*24] Saito S., Schwartz G., Simon T., Li J., Nam G.: Relightable gaussian codec avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 130–141. 1, 7, 8, 9, 27, 28

[SSSJ20] Saito S., Simon T., Saragih J., Joo H.: 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. 17, 21

[SSW*23] Shen Y., Saito S., Wang Z., Maury O., Wu C., Hodgins J., Zheng Y., Nam G.: Ct2hair: High-fidelity 3d hair modeling using computed tomography. *Transactions on Graphics, (Proc. SIGGRAPH) 42*, 4 (2023), 1–13. 12, 13, 32

[SSX*25] Sarafianos N., Stuyck T., Xiang X., Li Y., Popovic J., Ranjan R.: Garment3dgen: 3d garment stylization and texture generation. In *International Conference on 3D Vision (3DV)* (2025), IEEE, pp. 1382–1393. 17

[STH*24] Ströter D., Thiery J. M., Hormann K., Chen J., Chang Q., Besler S., Mueller-Roemer J. S., Boubekeur T., Stork A., Fellner D. W.: A survey on cage-based deformation of 3d models. *Computer Graphics Forum 43*, 2 (2024), e15060. 4

[SWL*24] Shao Z., Wang Z., Li Z., Wang D., Lin X., Zhang Y., Fan M., Wang Z.: Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 1606–1616. 9, 20

[SWW*20] Smith B., Wu C., Wen H., Peluse P., Sheikh Y., Hodgins J. K., Shiratori T.: Constraining dense hand surface tracking with elasticity. *Transactions on Graphics (TOG) 39*, 6 (2020), 1–14. 15

[SWW*23] Sun J., Wang X., Wang L., Li X., Zhang Y., Zhang H., Liu Y.: Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 20991–21002. 8, 9, 10, 11

[SXS*25] Shao R., Xu Y., Shen Y., Yang C., Zheng Y., Chen C., Liu Y., Wetzstein G.: Isa4d: Interspatial attention for efficient 4d human video generation. *SIGGRAPH* (2025). 21

[SYH*17] Shu Z., Yumer E., Hadap S., Sunkavalli K., Shechtman E., Samaras D.: Neural face editing with intrinsic image disentangling. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 5541–5550. 28

[SYMB21] Saito S., Yang J., Ma Q., Black M. J.: Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 2886–2897. 18

[SYZR21] Su S.-Y., Yu F., Zollhöfer M., Rhodin H.: A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Advances in Neural Information Processing Systems (NIPS)* (2021). 18

[SZG*25] Sang S., Zhi T., Gu T., Liu J., Luo L.: Lynx: Towards high-fidelity personalized video generation, 2025. 11

[SZH*23] Sklyarova V., Zakharov E., Hilliges O., Black M. J., Thies J.: Text-conditioned generative model of 3d strand-based human hairstyles. *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 4703–4712. 1, 12, 13, 14

[SZP*25] Sklyarova V., Zakharov E., Prinzler M., Becherini G., Black M., Thies J.: Im2haircut: Single-view strand-based hair reconstruction for human avatars. In *International Conference on Computer Vision (ICCV)* (2025). 13

[SZPF16] Schönberger J. L., Zheng E., Pollefeys M., Frahm J.-M.: Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)* (2016). 4, 13

[SZZ*22] Shao R., Zhang H., Zhang H., Chen M., Cao Y.-P., Yu T., Liu Y.: Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 15872–15882. 19

[TCL23] TSENG J., CASTELLON R., LIU C. K.: EDGE: editable dance generation from music. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), IEEE, pp. 448–458. 6

[TDK*25] TANG J., DAVOLI D., KIRSCHSTEIN T., SCHONEVELD L., NIESSNER M.: Gaf: Gaussian avatar reconstruction from monocular videos via multi-view diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 5546–5558. 5, 10, 11

[TFT*20] TEWARI A., FRIED O., THIES J., SITZMANN V., LOMBARDI S., SUNKAVALLI K., MARTIN-BRUALLA R., SIMON T., SARAGIH J., NIESSNER M., ET AL.: State of the art on neural rendering. *Computer Graphics Forum (CGF) 39*, 2 (2020), 701–727. 2

[TKB*23] TRETSCHK E., KAIRANDA N., B R M., DABRAL R., KORTYLEWSKI A., EGGER B., HABERMANN M., FUA P., THEOBALT C., GOLYANIK V.: State of the art in dense monocular non-rigid 3d reconstruction. *Computer Graphics Forum (Eurographics State of the Art Reports)* (2023). 2

[TKG*24] TEOTIA K., KIM H., GARRIDO P., HABERMANN M., ELGHARIB M., THEOBALT C.: Gaussianheads: End-to-end learning of drivable gaussian head avatars from coarse-to-fine representations. *Transactions on Graphics, (Proc. SIGGRAPH) 43*, 6 (2024), 1–12. 6, 7, 8

[TPT16] TKACH A., PAULY M., TAGLIASACCHI A.: Sphere-meshes for real-time hand modeling and tracking. *Transactions on Graphics (TOG) 35*, 6 (2016), 1–11. 14

[TRM*25] TEOTIA K., RHODIN H., MENDIRATTA M., KIM H., HABERMANN M., THEOBALT C.: Audio-driven universal gaussian head avatars, 2025. 11

[TSD*24] TU H., SHAO R., DONG X., ZHENG S., ZHANG H., CHEN L., WANG M., LI W., MA S., ZHANG S., ET AL.: Tele-aloha: A low-budget and high-authenticity telepresence system using sparse rgb cameras. In *SIGGRAPH Conference Papers (SA)* (2024), pp. 145–162. 19

[TST*15] TAGLIASACCHI A., SCHRÖDER M., TKACH A., BOUAZIZ S., BOTSCH M., PAULY M.: Robust articulated-icp for real-time hand tracking. *Computer Graphics Forum (CGF) 34*, 5 (2015), 101–114. 14

[TSTPM21] TIWARI G., SARAFIANOS N., TUNG T., PONS-MOLL G.: Neural-gif: Neural generalized implicit functions for animating people in clothing. In *International Conference on Computer Vision (ICCV)* (2021), pp. 11708–11718. 18

[TTM*22] TEWARI A., THIES J., MILDENHALL B., SRINIVASAN P., TRETSCHK E., YIFAN W., LASSNER C., SITZMANN V., MARTIN-BRUALLA R., LOMBARDI S., ET AL.: Advances in neural rendering. *Computer Graphics Forum (CGF) 41*, 2 (2022), 703–735. 2, 4

[TTS*24] TAKIMOTO Y., TAKEHARA H., SATO H., ZHU Z., ZHENG B.: Dr.hair: Reconstructing scalp-connected hair strands without pre-training via differentiable rendering of line segments. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 12, 13, 28, 29

[TXT*25] TAN J., XIANG D., TULSIANI S., RAMANAN D., YANG G.: Dressrecon: Freeform 4d human reconstruction from monocular video. In *International Conference on 3D Vision (3DV)* (2025), pp. 250–260. 17

[TZH*24a] TRAN P., ZAKHAROV E., HO L.-N., KARMANOV A., BERMUDEZ VENEGAS A., GOLDWHITE M., AGARWAL A., HU L., TRAN A., LI H.: Voodoo xp: Expressive one-shot head reenactment for vr telepresence. *Transactions on Graphics, (Proc. SIGGRAPH) 43*, 6 (2024), 1–26. 5, 6, 8, 10

[TZH*24b] TRAN P., ZAKHAROV E., HO L.-N., TRAN A. T., HU L., LI H.: Voodoo 3d: Volumetric portrait disentanglement for one-shot 3d head reenactment. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 10336–10348. 10

[TZN19] THIES J., ZOLLHÖFER M., NIESSNER M.: Deferred neural rendering: Image synthesis using neural textures. *Transactions on Graphics (TOG) 38*, 4 (2019), 1–12. 20

[TZS*16] 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)* (2016). 9, 22, 27

[TZT*25] TAUBNER F., ZHANG R., TULI M., BAHMANI S., LINDELL D. B.: Mvp4d: Multi-view portrait video diffusion for animatable 4d avatars. In *SIGGRAPH Asia Conference Papers (SA)* (2025), pp. 1–11. 5

[TZTL25] TAUBNER F., ZHANG R., TULI M., LINDELL D. B.: Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 5318–5330. 5, 8, 10, 11

[unr] Epic Games unreal engine. Unreal Engine. Accessed: 2025-10-10. 28, 29

[VCH*21] VILLEGAS R., CEYLAN D., HERTZMANN A., YANG J., SAITO J.: Contact-aware retargeting of skinned motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 9720–9729. 23

[VCR*18] VAROL G., CEYLAN D., RUSSELL B., YANG J., YUMER E., LAPTEV I., SCHMID C.: Bodynet: Volumetric inference of 3d human body shapes. In *European Conference on Computer Vision (ECCV)* (2018), pp. 20–36. 21

[VGC24] VIDAURRE R., GARCES E., CASAS D.: Diffusedwrinkles: A diffusion-based model for data-driven garment animation. *British Machine Vision Conference (BMVC)* (2024). 30

[VRM*17] VAROL G., ROMERO J., MARTIN X., MAHMOOD N., BLACK M. J., LAPTEV I., SCHMID C.: Learning from synthetic humans. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 109–117. 32

[VSGC20] VIDAURRE R., SANTESTEBAN I., GARCES E., CASAS D.: Fully convolutional graph neural networks for parametric virtual try-on. *Computer Graphics Forum (CGF) 39*, 8 (2020), 145–156. 30

[WAGT24] WANG S., ANTIC B., GEIGER A., TANG S.: Intrinsicavatar: Physically based inverse rendering of dynamic humans from monocular videos via explicit ray tracing. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 1877–1888. 8, 19, 20

[War92] WARD G. J.: Measuring and modeling anisotropic reflection. In *Proceedings of the 19th annual conference on Computer graphics and interactive techniques* (1992), pp. 265–272. 6

[WCHW24] WANG R., CAO Y., HAN K., WONG K.-Y. K.: A survey on 3d human avatar modeling–from reconstruction to generation. *arXiv* (2024). 2

[WCK*25] WANG J., CHEN M., KARAEV N., VEDALDI A., RUPPRECHT C., NOVOTNY D.: Vggt: Visual geometry grounded transformer. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), Computer Vision Foundation / IEEE. 4, 10

[WCL*23] WANG T., CHEN J., LI D., LIU X., WANG H., ZHOU K.: Fast gpu-based two-way continuous collision handling. *Transactions on Graphics (TOG) 42*, 5 (2023), 1–15. 30

[WCY*22] WANG L., CHEN Z., YU T., MA C., LI L., LIU Y.: Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). 32

[WJMLH01] WANN JENSEN H., MARSCHNER S. R., LEVOY M., HANRAHAN P.: A practical model for subsurface light transport. In *SIGGRAPH* (2001), ACM, pp. 511–518. 6

[WLC*24a] WANG S., LEROY V., CABON Y., CHIDLOVSKII B., REVAUD J.: Dust3r: Geometric 3d vision made easy. In *CVPR* (2024). 4

[WLC*24b] WANG S., LEROY V., CABON Y., CHIDLOVSKII B., REVAUD J.: Dust3r: Geometric 3d vision made easy. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), Computer Vision Foundation / IEEE, pp. 20697–20709. 10

[WLD*24] WANG J., LIU Y., DOU Z., YU Z., LIANG Y., LIN C., XIE R., SONG L., LI X., WANG W.: Disentangled clothed avatar generation

from text descriptions. In *European Conference on Computer Vision (ECCV)* (2024), pp. 381–401. 18

[WLZ*25] WANG Z., LI Y., ZENG Y., GUO Y., LIN D., XUE T., DAI B.: Multi-identity human image animation with structural video diffusion. *arXiv* (2025). 21

[WMB19] WANG B., MATCUK G., BARBIČ J.: Hand modeling and simulation using stabilized magnetic resonance imaging. *Transactions on Graphics (TOG) 38*, 4 (2019), 1–14. 15

[WML21] WANG T.-C., MALLYA A., LIU M.-Y.: One-shot free-view neural talking-head synthesis for video conferencing. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 10039–10049. 11

[WMM*21] WANG S., MIHAJLOVIC M., MA Q., GEIGER A., TANG S.: Metaavatar: Learning animatable clothed human models from few depth images. *Advances in Neural Information Processing Systems (NeurIPS) 34* (2021), 2810–2822. 20

[WNS*22] WANG Z., NAM G., STUYCK T., LOMBARDI S., ZOLLHOEFER M., HODGINS J., LASSNER C.: Hvh: Learning a hybrid neural volumetric representation for dynamic hair performance capture. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). 28, 29

[WNS*23] WANG Z., NAM G., STUYCK T., LOMBARDI S., CAO C., SARAGIH J., ZOLLHOEFER M., HODGINS J., LASSNER C.: Neuwigs: A neural dynamic model for volumetric hair capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 28, 29, 30

[Woo80] WOODHAM R. J.: Photometric method for determining surface orientation from multiple images. *Optical engineering 19*, 1 (1980), 139–144. 16

[WOR11] WANG H., O'BRIEN J. F., RAMAMOORTHI R.: Data-driven elastic models for cloth: modeling and measurement. *Transactions on Graphics, (Proc. SIGGRAPH) 30*, 4 (2011), 1–12. 17

[WPJT25] WU Y., PRINZLER M., JIN X., TANG S.: Text-based animatable 3d avatars with morphable model alignment. In *SIGGRAPH Conference Papers (SA)* (2025), pp. 1–11. 5, 8, 11

[WPW*25] WANG R., PRADA F., WANG Z., JIANG Z., YIN C., LI J., SAITO S., SANTESTEBAN I., ROMERO J., JOSHI R., ET AL.: Fresa: Feedforward reconstruction of personalized skinned avatars from few images. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025). 8, 20

[WRG*09] WANG J., REN P., GONG M., SNYDER J., GUO B.: All-frequency rendering of dynamic, spatially-varying reflectance. *Transactions on Graphics, (Proc. SIGGRAPH Asia) 28*, 5 (2009), 133. 28

[WS00] WYSZECKI G., STILES W. S.: *Color science: concepts and methods, quantitative data and formulae*. John wiley & sons, 2000. 27

[WSCKS23] WENG C.-Y., SRINIVASAN P. P., CURLESS B., KEMELMACHER-SHLIZERMAN I.: Personnerf: Personalized reconstruction from photo collections. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 524–533. 19, 20

[WSS*25] WANG S., SIMON T., SANTESTEBAN I., BAGAUTDINOV T., LI J., AGRAWAL V., PRADA F., YU S.-I., NALBONE P., GRAMLICH M., ET AL.: Relightable full-body gaussian codec avatars. In *SIGGRAPH Conference Papers (SA)* (2025), pp. 1–12. 8, 19, 27, 28

[WTD*25] WANG H., TANG H., DI D., ZHANG Z., ZUO W., GAO F., MA S., ZHANG S.: Mosa: Motion-coherent human video generation via structure-appearance decoupling. *arXiv* (2025). 21

[WWA*25] WAN T., WANG A., AI B., WEN B., MAO C., XIE C.-W., CHEN D., YU F., ZHAO H., YANG J., ET AL.: Wan: Open and advanced large-scale video generative models. *arXiv* (2025). 5, 14

[WWJ*25] WANG M., WANG Q., JIANG F., FAN Y., ZHANG Y., QI Y., ZHAO K., XU M.: Fantasytalking: Realistic talking portrait generation via coherent motion synthesis, 2025. 11

[WWO*19] WANG S.-Y., WANG O., OWENS A., ZHANG R., EFROS A. A.: Detecting photoshopped faces by scripting photoshop. In *International Conference on Computer Vision (ICCV)* (2019), pp. 10072–10081. 22

[WWS*20] WANG K., WU Q., SONG L., YANG Z., WU W., QIAN C., HE R., QIAO Y., LOY C. C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision (ECCV)* (2020), Springer, pp. 700–717. 32

[WXT*25] WANG J., XU Y., TRETSCHK E., WANG Z., IANINA A., BOZIC A., NEUMANN U., TUNG T.: Dgh: Dynamic gaussian hair. In *Advances in Neural Information Processing Systems (NeurIPS)* (2025). 29, 30

[WYK*24] WU K., YANG L., KUANG Z., FENG Y., HAN X., SHEN Y., FU H., ZHOU K., ZHENG Y.: Monohair: High-fidelity hair modeling from a monocular video. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 12

[WYL*25] WEI J., YAN S., LIN W., LIU B., CHEN R., GUO M.: Echovideo: Identity-preserving human video generation by multimodal feature fusion. *arXiv* (2025). 21

[WYY*22] WU K., YE Y., YANG L., FU H., ZHOU K., ZHENGL Y.: NeuralHDHair: Automatic high-fidelity hair modeling from a single image using implicit neural representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 1516–1525. 12, 13

[WZG*26] WANG L., ZHU Y., GE Z., ZHENG Y., ZHANG L., HU T., QIN S., LUO M., ZHANG J., CHEN X., ET AL.: Flowact-r1: Towards interactive humanoid video generation. *arXiv* (2026). 21

[WZL*18] WANG B., ZHENG H., LIANG X., CHEN Y., LIN L., YANG M.: Toward characteristic-preserving image-based virtual try-on network. In *European Conference on Computer Vision (ECCV)* (2018), pp. 589–604. 31

[WZN*25] WANG C., ZHUANG P., NGO T. D., MENAPACE W., SIAROHIN A., VASILKOVSKY M., SKOROKHODOV I., TULYAKOV S., WONKA P., LEE H.-Y.: 4real-video: Learning generalizable photorealistic 4d video diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 17723–17732. 21

[WZR*24] WEN J., ZHAO X., REN Z., SCHWING A. G., WANG S.: Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 2059–2069. 20

[XAS21] XU H., ALLDIECK T., SMINCHISESCU C.: H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems (NeurIPS) 34* (2021), 14955–14966. 20

[XBS*22] XIANG D., BAGAUTDINOV T., STUYCK T., PRADA F., ROMERO J., XU W., SAITO S., GUO J., SMITH B., SHIRATORI T., ET AL.: Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *Transactions on Graphics, (Proc. SIGGRAPH Asia) 41*, 6 (2022), 1–15. 6, 16

[XCG*24] XU S., CHEN G., GUO Y., YANG J., LI C., ZANG Z., ZHANG Y., TONG X., GUO B.: VASA-1: lifelike audio-driven talking faces generated in real time. In *Advances in Neural Information Processing Systems (NeurIPS)* (2024). 2, 6, 8, 11

[XCL*24] XU Y., CHEN B., LI Z., ZHANG H., WANG L., ZHENG Z., LIU Y.: Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 1931–1941. 7

[XGGZ24] XIANG J., GAO X., GUO Y., ZHANG J.: Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 1802–1812. 9

[XGL*25] XIE Y., GU T., LI Z., ZHANG C., SONG G., ZHAO X., LIANG C., JIANG J., XU H., LUO L.: X-streamer: Unified human world modeling with audiovisual interaction. *arXiv* (2025). 21

[XMR*11] XU K., MA L.-Q., REN B., WANG R., HU S.-M.: Interactive hair rendering and appearance editing under environment lighting. *Transactions on Graphics (TOG) 30*, 6 (2011), 173:1–173:10. 28

[XPB*21] XIANG D., PRADA F., BAGAUTDINOV T., XU W., DONG Y., WEN H., HODGINS J., WU C.: Modeling clothing as a separate layer for an animatable human avatar. *Transactions on Graphics, (Proc. SIGGRAPH Asia) 40*, 6 (2021), 1–15. 16

[XPC*23] XIANG D., PRADA F., CAO Z., GUO K., WU C., HODGINS J., BAGAUTDINOV T.: Drivable avatar clothing: Faithful full-body telepresence with dynamic clothing driven by sparse rgb-d input. In *SIGGRAPH Asia Conference Papers (SA)* (2023), pp. 1–11. 16

[XPG*24] XU Z., PENG S., GENG C., MOU L., YAN Z., SUN J., BAO H., ZHOU X.: Relightable and animatable neural avatar from sparse-view video. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 990–1000. 8, 19

[XPWH20] XIANG D., PRADA F., WU C., HODGINS J.: Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *International Conference on 3D Vision (3DV)* (2020), pp. 322–332. 17

[XTS*22] XIE Y., TAKIKAWA T., SAITO S., LITANY O., YAN S., KHAN N., TOMBARI F., TOMPKIN J., SITZMANN V., SRIDHAR S.: Neural fields in visual computing and beyond. *Computer Graphics Forum (CGF) 41*, 2 (2022), 641–676. 15

[XWW*14] XU Z., WU H.-T., WANG L., ZHENG C., TONG X., QI Y.: Dynamic hair capture using spacetime optimization. *Transactions on Graphics, (Proc. SIGGRAPH) 33*, 6 (2014). 29

[XWZ*22] XIE L., WANG X., ZHANG H., DONG C., SHAN Y.: Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 657–666. 24, 32

[XWZ*24] XU Y., WANG L., ZHENG Z., SU Z., LIU Y.: 3d gaussian parametric head model. In *European Conference on Computer Vision (ECCV)* (2024), Springer, pp. 129–147. 8, 9, 10

[XYC*23] XIU Y., YANG J., CAO X., TZIONAS D., BLACK M. J.: Econ: Explicit clothed humans optimized via normal integration. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 512–523. 21

[XYL*24] XIU Y., YE Y., LIU Z., TZIONAS D., BLACK M. J.: Puzzleavatar: Assembling 3d avatars from personal albums. *Transactions on Graphics, (Proc. SIGGRAPH) 43*, 6 (2024), 1–15. 8, 20

[XYTB22] XIU Y., YANG J., TZIONAS D., BLACK M. J.: Icon: Implicit clothed humans obtained from normals. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 13286–13296. 8, 21

[XZJJ25] XIE T., ZHAO Y., JIANG Y., JIANG C.: Physanimator: Physics-guided generative cartoon animation, 2025. 14

[XZL*23] XU Z., ZHANG J., LIEW J. H., FENG J., SHOU M. Z.: XA-Gen: 3D expressive human avatars generation. In *Advances in Neural Information Processing Systems (NeurIPS)* (2023). 21

[YA06] YUKSEL C., AKLEMAN E.: Rendering hair with global illumination. In *ACM SIGGRAPH 2006 Research Posters* (New York, NY, USA, 2006), ACM Press, p. 124. 28

[YBM*24] YU Z., BAI Z., MEKA A., TAN F., XU Q., PANDEY R., FANELLO S., PARK H. S., ZHANG Y.: One2avatar: Generative implicit head avatar for few-shot user adaptation. *arXiv* (2024). 10

[YCG*24] YANG X., CHEN X., GAO D., WANG S., HAN X., WANG B.: Have-fun: Human avatar reconstruction from few-shot unconstrained images. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 742–752. 20

[YCW*22] YOON J. S., CEYLAN D., WANG T. Y., LU J., YANG J., SHU Z., PARK H. S.: Learning motion-dependent appearance for high-fidelity rendering of dynamic humans from a single camera. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 3407–3417. 20

[YFZ*23] YU W., FAN Y., ZHANG Y., WANG X., YIN F., BAI Y., CAO Y.-P., SHAN Y., WU Y., SUN Z., ET AL.: Nofa: Nerf-based one-shot facial avatar reconstruction. In *SIGGRAPH Conference Papers (SA)* (2023), pp. 1–12. 10

[YK08] YUKSEL C., KEYSER J.: Deep opacity maps. *Computer Graphics Forum (CGF) 27*, 2 (2008), 675–680. 28

[YLB*25] YU Z., LI Z., BAO H., YANG C., ZHOU X.: Humanram: Feed-forward human reconstruction and animation model using transformers. In *SIGGRAPH Conference Papers (SA)* (2025). 6, 8

[YLL*25] YANG Y., LIU F., LU Y., ZHAO Q., WU P., ZHAI W., YI R., CAO Y., MA L., ZHA Z.-J., ET AL.: Sigman: Scaling 3d human gaussian generation with millions of assets. In *International Conference on Computer Vision (ICCV)* (2025), pp. 1–11. 21

[YLS*25] YU Z., LI T., SUN J., SHAPIRA O., PARK S., STENGEL M., CHAN M., LI X., WANG W., NAGANO K., ET AL.: Gaia: Generative animatable interactive avatars with expression-conditioned gaussians. In *SIGGRAPH Conference Papers (SA)* (2025), pp. 1–10. 11

[YSC*25] YANG Y., SHENG H., CAI S., LIN J., WANG J., DENG B., LU J., WANG H., YE J.: Echomotion: Unified human video and motion generation via dual-modality diffusion transformer. *arXiv* (2025). 21

[YSZZ19] YANG L., SHI Z., ZHENG Y., ZHOU K.: Dynamic hair modeling from monocular videos using deep neural networks. *Transactions on Graphics, (Proc. SIGGRAPH) 38*, 6 (2019). 29

[YTB*21] YENAMANDRA T., TEWARI A., BERNARD F., SEIDEL H.-P., ELGHARIB M., CREMERS D., THEOBALT C.: i3dmm: Deep implicit 3d morphable model of human heads. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 12803–12813. 5

[YVN*22] YANG G., VO M., NEVEROVA N., RAMANAN D., VEDALDI A., JOO H.: Banmo: Building animatable 3d neural models from many casual videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 2863–2873. 17

[YWO09] YAMAGUCHI T., WILBURN B., OFEK E.: Video-based modeling of dynamic hair. In *Pacific-Rim Symposium on Image and Video Technology (PSIVT)* (2009), pp. 585–596. 29

[YZG*21] YU T., ZHENG Z., GUO K., LIU P., DAI Q., LIU Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 5746–5756. 3, 32

[YZM*24] YANG H., ZHENG M., MA C., LAI Y.-K., WAN P., HUANG H.: Vrmm: A volumetric relightable morphable head model. In *SIGGRAPH Conference Papers (SA)* (2024), pp. 1–11. 10

[YZR*24] YE Z., ZHONG T., REN Y., YANG J., LI W., HUANG J., JIANG Z., HE J., HUANG R., LIU J., ET AL.: Real3d-portrait: One-shot realistic 3d talking portrait synthesis. In *International Conference on Learning Representations (ICLR)* (2024). 5, 10

[ZAB*22] ZHENG Y., ABREVAYA V. F., BÜHLER M. C., CHEN X., BLACK M. J., HILLIGES O.: Im avatar: Implicit morphable head avatars from videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 13545–13555. 9

[ZBBT25] ZIELONKA W., BOLKART T., BEELER T., THIES J.: Gaussian eigen models for human heads. *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025). 6, 7

[ZBS*25] ZIELONKA W., BAGAUTDINOV T., SAITO S., ZOLLHÖFER M., THIES J., ROMERO J.: Drivable 3d gaussian avatars. *International Conference on 3D Vision (3DV)* (2025). 8, 16, 19

[ZBT22] ZIELONKA W., BOLKART T., THIES J.: Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 250–269. 27

[ZBT23] ZIELONKA W., BOLKART T., THIES J.: Instant volumetric head avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 4574–4584. 7, 8, 9

[ZBT*24] ZHANG K., BI S., TAN H., XIANGLI Y., ZHAO N., SUNKAVALLI K., XU Z.: Gs-lrm: Large reconstruction model for 3d

gaussian splatting. In *European Conference on Computer Vision (ECCV)* (2024), Springer, pp. 1–19. 10

[ZCD*24] ZHU S., CHEN J. L., DAI Z., DONG Z., XU Y., CAO X., YAO Y., ZHU H., ZHU S.: Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision (ECCV)* (2024), pp. 145–162. 21

[ZCM22] ZHANG M., CEYLAN D., MITRA N. J.: Motion guided deep dynamic 3d garments. *Transactions on Graphics, (Proc. SIGGRAPH Asia) 41*, 6 (2022), 1–12. 30

[ZCP*23] ZHOU Y., CHAI M., PEPE A., GROSS M., BEELER T.: Groomgen: A high-quality generative hair model using hierarchical latent representations. *Transactions on Graphics, (Proc. SIGGRAPH)* (2023). 12, 13, 29, 30

[ZCV*25] ZHENG Y., CHAI M., VICINI D., ZHOU Y., XU Y., GUIBAS L., WETZSTEIN G., BEELER T.: Groomlight: Hybrid inverse rendering for relightable human hair appearance modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 16040–16050. 12, 28, 29

[ZCW*17] ZHANG M., CHAI M., WU H., YANG H., ZHOU K.: A data-driven approach to four-view image-based hair modeling. *Transactions on Graphics, (Proc. SIGGRAPH) 36*, 4 (2017). 12

[ZCW*24a] ZHANG B., CHENG Y., WANG C., ZHANG T., YANG J., TANG Y., ZHAO F., CHEN D., GUO B.: Rodinhd: High-fidelity 3d avatar generation with diffusion models. In *European Conference on Computer Vision (ECCV)* (2024), Springer, pp. 465–483. 11

[ZCW*24b] ZHOU Y., CHAI M., WANG D., WINBERG S., WOOD E., SARKAR K., GROSS M., BEELER T.: Groomcap: High-fidelity prior-free hair capture. *Transactions on Graphics, (Proc. SIGGRAPH)* (2024). 12, 13, 28, 29

[ZCW*25] ZHAI Z., CHEN G., WANG W., ZHENG D., XIAO J.: Taga: Self-supervised learning for template-free animatable gaussian articulated model. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 21159–21169. 20

[ZFK*24] ZHANG H., FENG Y., KULITS P., WEN Y., THIES J., BLACK M. J.: Teca: Text-guided generation and editing of compositional 3d avatars. In *International Conference on 3D Vision (3DV)* (2024), IEEE, pp. 1520–1530. 11

[ZGL*25] ZIELONKA W., GARBIN S. J., LATTAS A., KOPANAS G., GOTARDO P., BEELER T., THIES J., BOLKART T.: Synthetic prior for few-shot drivable head avatar inversion. *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025). 9, 23

[ZGW*25] ZHANG Y., GU J., WANG L.-W., WANG H., ZHU Y., ZOU F., ET AL.: Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. In *International Conference on Machine Learning (ICML)* (2025), pp. 1–10. 21

[ZHX*18] ZHOU Y., HU L., XING J., CHEN W., KUNG H.-W., TONG X., LI H.: HairNet: Single-view hair reconstruction using convolutional neural networks. In *European Conference on Computer Vision (ECCV)* (2018), pp. 249–265. 12, 13

[ZHY*22] ZHENG Z., HUANG H., YU T., ZHANG H., GUO Y., LIU Y.: Structured local radiance fields for human avatar modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 15893–15903. 19

[ZJL*23] ZHENG Y., JIN Z., LI M., HUANG H., MA C., CUI S., HAN X.: Hairstep: Transfer synthetic to real using strand and depth maps for single-view 3d hair modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 12726–12735. 12, 13, 14

[ZKB*25] ZHUANG J., KANG D., BAO L., LIN L., LI G.: Dagsm: Disentangled avatar generation with gs-enhanced mesh. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 292–303. 18

[ZLB*24] ZHANG J. X., LIN G. W.-C., BODE L., CHEN H.-Y., STUYCK T., LARIONOV E.: Estimating cloth elasticity parameters from

homogenized yarn-level models. In *Proceedings of the 17th ACM SIGGRAPH Conference on Motion, Interaction, and Games* (2024), pp. 1–12. 17

[ZLDF21] ZHANG Z., LI L., DING Y., FAN C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 3661–3670. 24

[ZLL*24] ZHU L., LI Y., LIU N., PENG H., YANG D., KEMELMACHER-SHLIZERMAN I.: M&m vto: Multi-garment virtual try-on and editing. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 1346–1356. 31

[ZLS*23] ZHANG H., LIN S., SHAO R., ZHANG Y., ZHENG Z., HUANG H., GUO Y., LIU Y.: Closet: Modeling clothed humans on continuous surface with explicit template decomposition. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 501–511. 18

[ZLW*21] ZHANG K., LUAN F., WANG Q., BALA K., SNAVELY N.: Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), Computer Vision Foundation / IEEE, pp. 5453–5462. 28

[ZLW*25] ZHUANG Y., LV J., WEN H., SHUAI Q., ZENG A., ZHU H., CHEN S., YANG Y., CAO X., LIU W.: Idol: Instant photorealistic 3d human creation from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 26308–26319. 8, 21

[ZMF*24] ZHOU Z., MA F., FAN H., YANG Z., YANG Y.: Headstudio: Text to animatable head avatars with 3d gaussian splatting. In *European Conference on Computer Vision (ECCV)* (2024). 5, 8, 9, 11

[ZMFC25] ZHOU Z., MA F., FAN H., CHUA T.-S.: Zero-1-to-a: Zero-shot one image to animatable head avatars using video diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 15941–15952. 5, 10

[ZMZ*23] ZHANG S., MA Q., ZHANG Y., ALIAKBARIAN S., COSKER D., TANG S.: Probabilistic human mesh recovery in 3d scenes from egocentric views. In *International Conference on Computer Vision (ICCV)* (2023), pp. 7989–8000. 21

[ZPBPM17] ZHANG C., PUJADES S., BLACK M. J., PONS-MOLL G.: Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 4191–4200. 16

[ZPX*24] ZHOU X., PENG S., XU Z., DONG J., WANG Q., ZHANG S., SHUAI Q., BAO H.: Animatable implicit neural representations for creating realistic avatars from videos. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 46*, 6 (2024), 4147–4159. 18

[ZQQH22] ZHU H., QIU L., QIU Y., HAN X.: Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 3845–3854. 17

[ZRA23] ZHANG L., RAO A., AGRAWALA M.: Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision (ICCV)* (2023), pp. 3836–3847. 5, 14, 21

[ZRL*09] ZINKE A., RUMP M., LAY T., WEBER A., ANDRIYENKO A., KLEIN R.: A practical approach for photometric acquisition of hair color. *Transactions on Graphics, (Proc. SIGGRAPH 28*, 5 (2009), 1–9. 28

[ZSB*24] ZAKHAROV E., SKLYAROVA V., BLACK M., NAM G., THIES J., HILLIGES O.: Human hair reconstruction with strand-aligned 3d gaussians. In *European Conference on Computer Vision (ECCV)* (2024), Springer, pp. 409–425. 6, 12, 13, 28, 29

[ZSW*24] ZHAO X., SUN J., WANG L., SUO J., LIU Y.: Invertavatar: Incremental gan inversion for generalized head avatars. In *SIGGRAPH Conference Papers (SA)* (2024), pp. 1–10. 8

[ZSYZ25] ZHAN Y., SHAO T., YANG Y., ZHOU K.: Real-time high-fidelity gaussian human avatars with position-based interpolation of spatially distributed mlps. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 26297–26307. 19

[ZTG*18] 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 (CGF) 37*, 2 (2018), 523–550. 2, 4

[ZWHB22] ZHENG M., WANG B., HUANG J., BARBIČ J.: Simulation of hand anatomy using medical imaging. *Transactions on Graphics (TOG) 41*, 6 (2022), 1–20. 15

[ZWL*22] ZHUO L., WANG G., LI S., WU W., LIU Z.: Fast-vid2vid: Spatial-temporal compression for video-to-video synthesis. In *European Conference on Computer Vision (ECCV)* (2022), pp. 289–305. 20

[ZWL*25] ZHENG X., WEN C., LI Z., ZHANG W., SU Z., CHANG X., ZHAO Y., LV Z., ZHANG X., ZHANG Y., ET AL.: Headgap: Few-shot 3d head avatar via generalizable gaussian priors. In *International Conference on 3D Vision (3DV)* (2025), IEEE, pp. 946–957. 10

[ZWLY25] ZHANG W., WU S., LIAO M., YAN Y.: Disentangled clothed avatar generation with layered representation. In *International Conference on Computer Vision (ICCV)* (2025). 18

[ZWS*24] ZHENG X., WEN C., SU Z., XU Z., LI Z., ZHAO Y., XUE Z.: Ohta: One-shot hand avatar via data-driven implicit priors. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 799–810. 15

[ZWW*18] ZHANG M., WU P., WU H., WENG Y., ZHENG Y., ZHOU K.: Modeling hair from an rgb-d camera. *Transactions on Graphics, (Proc. SIGGRAPH) 37*, 6 (2018), 205:1–205:10. 12, 13

[ZWZ*22] ZHU H., WU W., ZHU W., JIANG L., TANG S., ZHANG L., LIU Z., LOY C. C.: Celebv-hq: A large-scale video facial attributes dataset. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 650–667. 32

[ZYG*23] ZHU H., YANG H., GUO L., ZHANG Y., WANG Y., HUANG M., WU M., SHEN Q., YANG R., CAO X.: Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 45*, 12 (2023), 14528–14545. 32

[ZYHC22] ZHENG M., YANG H., HUANG D., CHEN L.: Imface: A nonlinear 3d morphable face model with implicit neural representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 20343–20352. 5

[ZYL*24] ZHANG W., YAN Y., LIU Y., SHENG X., YANG X.: E^3gen: Efficient, expressive and editable avatars generation. In *International Conference on Multimedia, MM* (2024), ACM, pp. 6860–6869. 21

[ZYW*23] ZHENG Y., YIFAN W., WETZSTEIN G., BLACK M. J., HILLIGES O.: Pointavatar: Deformable point-based head avatars from videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 21057–21067. 9

[ZYWK08] ZINKE A., YUKSEL C., WEBER A., KEYSER J.: Dual scattering approximation for fast multiple scattering in hair. *Transactions on Graphics, (Proc. SIGGRAPH) 27*, 3 (2008), 32:1–32:10. 6, 28

[ZYX*24] ZHANG Y., YE D., XIE C., TANG L., LIAO X., LIU Z., CHEN C., DENG J.: Dual defense: Adversarial, traceable, and invisible robust watermarking against face swapping. *IEEE Transactions on Information Forensics and Security 19* (2024), 4628–4641. 22

[ZYZ*23] ZHU L., YANG D., ZHU T., REDA F., CHAN W., SAHARIA C., NOROUZI M., KEMELMACHER-SHLIZERMAN I.: Tryondiffusion: A tale of two unets. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 4606–4615. 31

[ZZ19] ZHANG M., ZHENG Y.: Hair-gan: Recovering 3d hair structure from a single image using generative adversarial networks. *Visual Informatics (VI) 3*, 2 (2019), 102–112. 12, 13

[ZZC*23] ZHANG X., ZHANG J., CHACKO R., XU H., SONG G., YANG Y., FENG J.: Getavatar: Generative textured meshes for animatable human avatars. In *International Conference on Computer Vision (ICCV)* (2023), pp. 2273–2282. 18

[ZZR*25] ZHU Y., ZHANG L., RONG Z., HU T., LIANG S., GE Z.: Infp: Audio-driven interactive head generation in dyadic conversations. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (2025), pp. 10667–10677. 23

[ZZS*24] ZHENG S., ZHOU B., SHAO R., LIU B., ZHANG S., NIE L., LIU Y.: 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. 8

[ZZSC22] ZHUANG Y., ZHU H., SUN X., CAO X.: Mofanerf: Morphable facial neural radiance field. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 268–285. 10

[ZZTH24] ZHU H., ZHAN F., THEOBALT C., HABERMANN M.: Tri-human: a real-time and controllable tri-plane representation for detailed human geometry and appearance synthesis. *Transactions on Graphics, (Proc. SIGGRAPH) 44*, 1 (2024), 1–17. 19

[ZZY*24] ZHENG Y., ZHAO Q., YANG G., YIFAN W., XIANG D., DUBOST F., LAGUN D., BEELER T., TOMBARI F., GUIBAS L., ET AL.: Physavatar: Learning the physics of dressed 3d avatars from visual observations. In *European Conference on Computer Vision (ECCV)* (2024), Springer, pp. 262–284. 6, 16, 17, 19

[ZZZ*23] ZHENG Z., ZHAO X., ZHANG H., LIU B., LIU Y.: Avatarrex: Real-time expressive full-body avatars. *Transactions on Graphics, (Proc. SIGGRAPH) 42*, 4 (2023), 1–19. 8, 19