Capturing and Synthesis of 3D Digital Humans

Wojciech Zielonka

Committee:

Prof. Justus Thies
Prof. Matthias Niessner
Prof. Jan Gugenheimer
Prof. Simone Schaub-Meyer

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Meta Codec Avatars

CODEC AVATAR

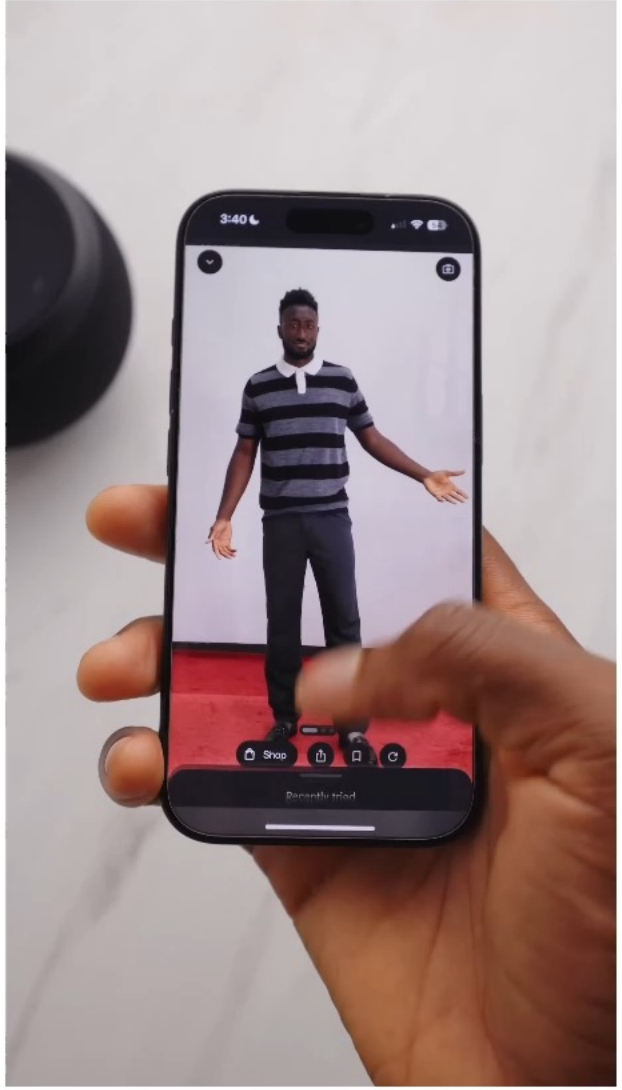

Meta Codec Avatars

Holoportation Microsoft Research

Potential Applications of Digital Avatars

🧠 Research & Simulation [1]

🛍️ Retail & Fashion [2]

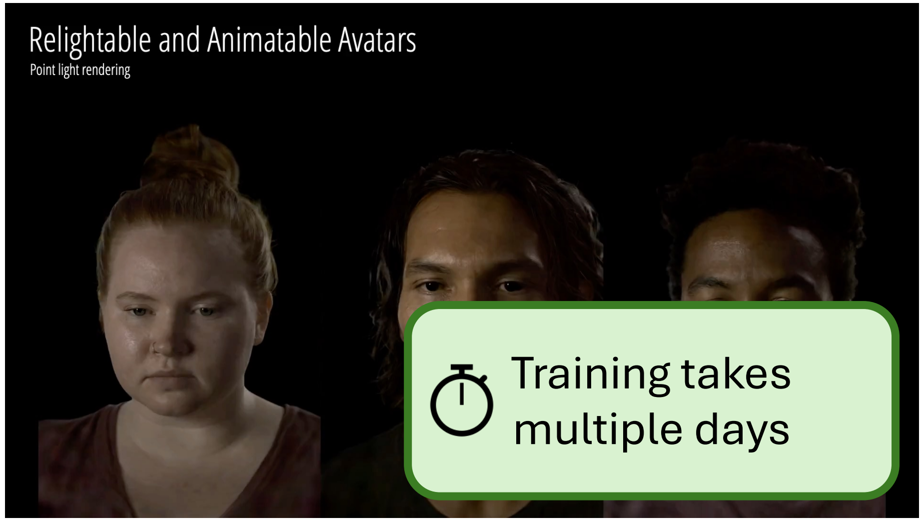


🤖 Human-Computer Interaction [3]

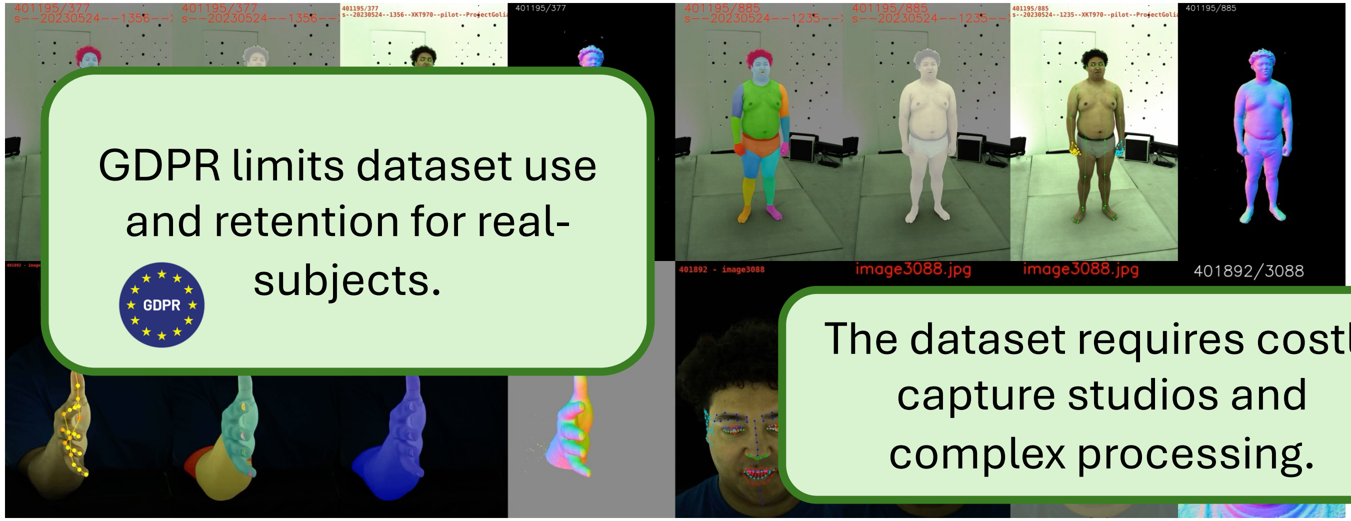
🎮 Entertainment & Media [4]

1) Human behavior modeling
2) Google Virtual Try-on
3) Apple Persona
4) Runways ACT 2

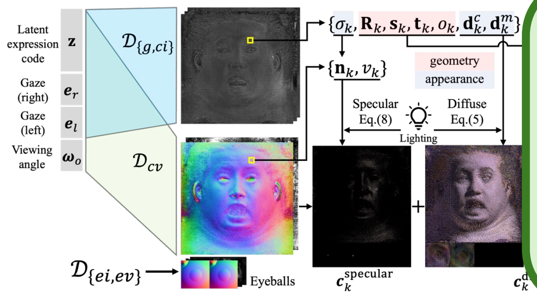
Why is it a Hard Problem?

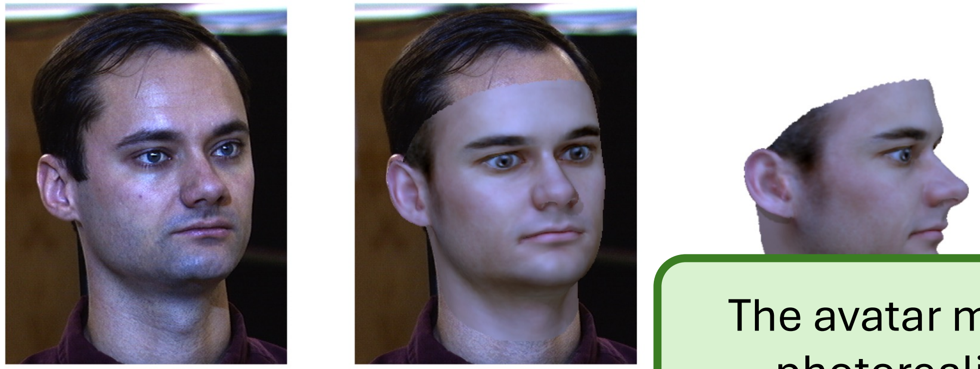


Photorealism comes at the cost of sophisticated and powerful networks. [**2**]



Creating photorealistic 3D digital avatars requires assets that take up hundreds of gigabytes and take weeks to process. [**3**]

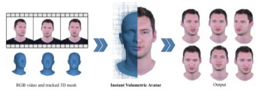




Any lack of photorealism immediat[...] immersion and makes the avatars look artificial. [**1**]

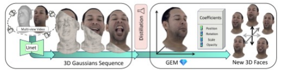
1) A 3D Face Model for Pose and Illumination Invariant Face Recognition, Paysan *et al.*
2) Relightable Gaussian Codec Avatars, Saito *et al.*
3) Codec Avatar Studio, Martinez *et al.*

Challenges Addressed in this PhD Work



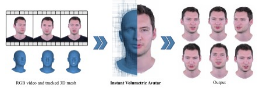
☑ Speed of the avatar's creation.

☑ Few-shot inversion by utilizing synthetic prior.



☑ Distillation to lightweight representation.

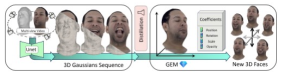
Challenges Addressed in this PhD Work



☑ Speed of the avatar's creation.

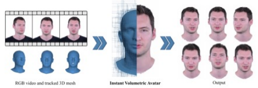
- Many methods need days (even up to 5 days) to train a single avatar.
- Appearance changes were not address due to compute heavy training.
- They often lack real-time capabilities.

☑ Few-shot inversion by utilizing synthetic prior.



☑ Distillation to lightweight representation.

Challenges Addressed in this PhD Work

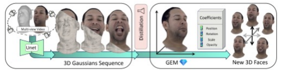


☑ Speed of the avatar's creation.

- Many methods need days (even up to 5 days) to train a single avatar.
- Appearance changes were not address due to compute heavy training.
- They often lack real-time capabilities.

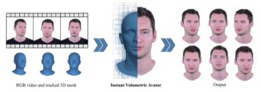
☑ Few-shot inversion by utilizing synthetic prior.

- Monocular avatars don't generalize well to new views or expressions.
- 3D models need diverse paired data, which is scarce.
- GDPR limits dataset use and retention.



☑ Distillation to lightweight representation.

Challenges Addressed in this PhD Work

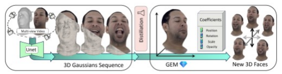


☑ Speed of the avatar's creation.

- Many methods need days (even up to 5 days) to train a single avatar.
- Appearance changes were not address due to compute heavy training.
- They often lack real-time capabilities.

☑ Few-shot inversion by utilizing synthetic prior.

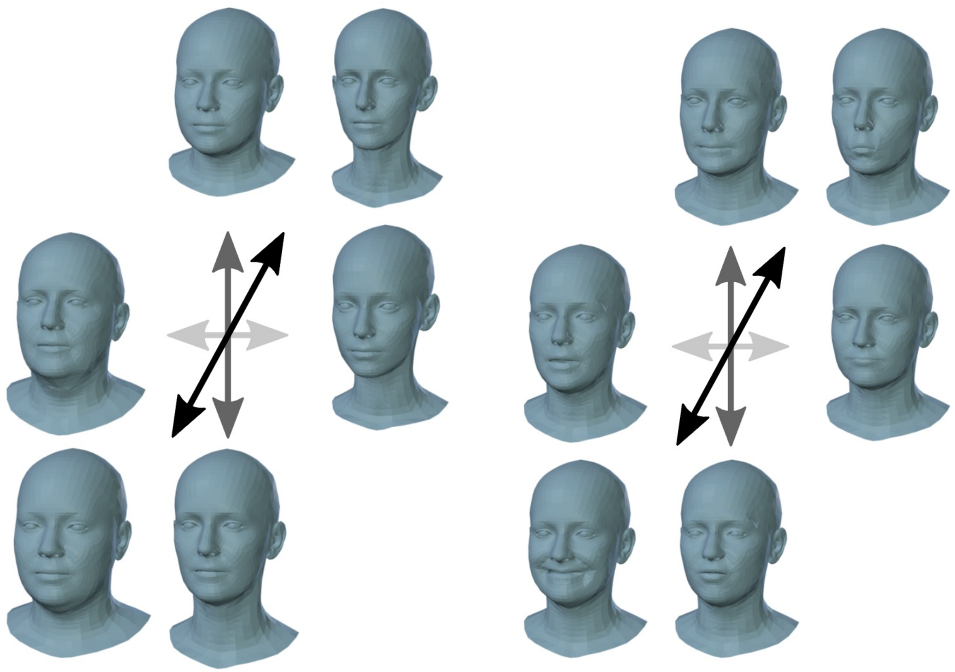
- Monocular avatars don't generalize well to new views or expressions.
- 3D models need diverse paired data, which is scarce.
- GDPR limits dataset use and retention.



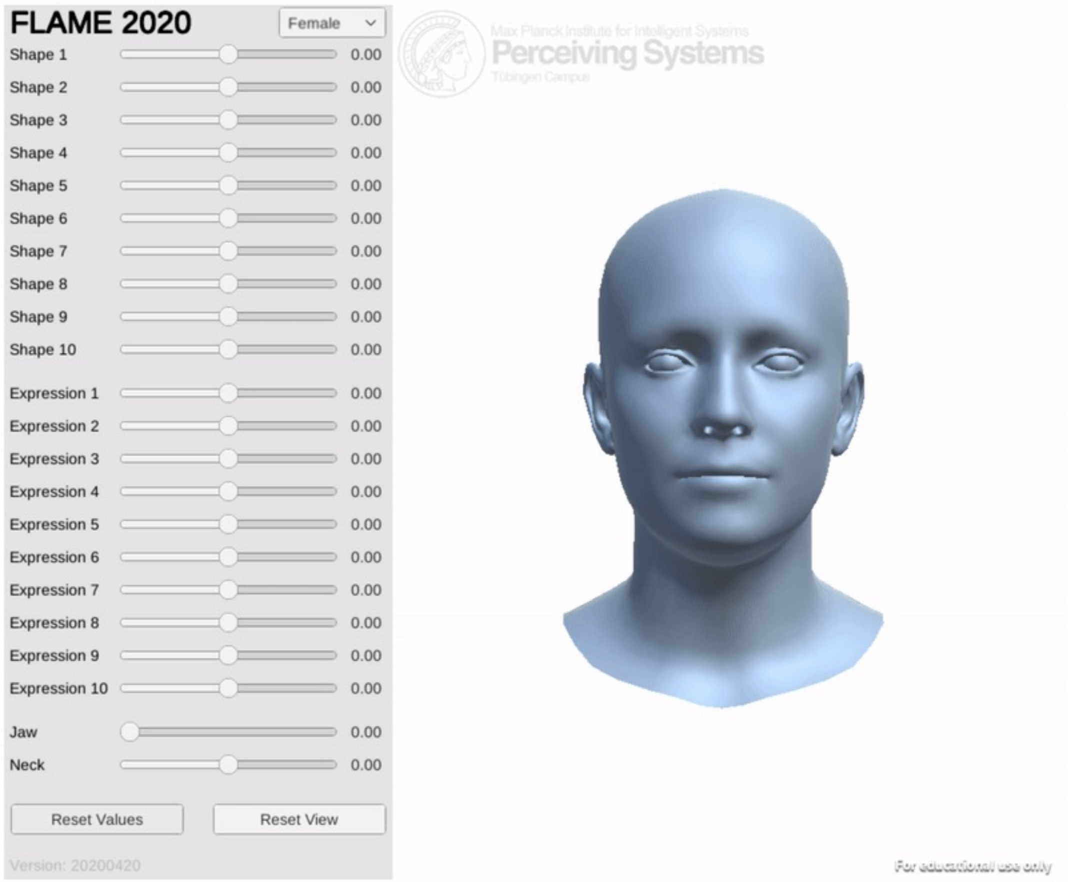
☑ Distillation to lightweight representation.

- Lightweight models look unrealistic.
- High-quality avatars are too resource-heavy for common devices.

FLAME*



shape expression



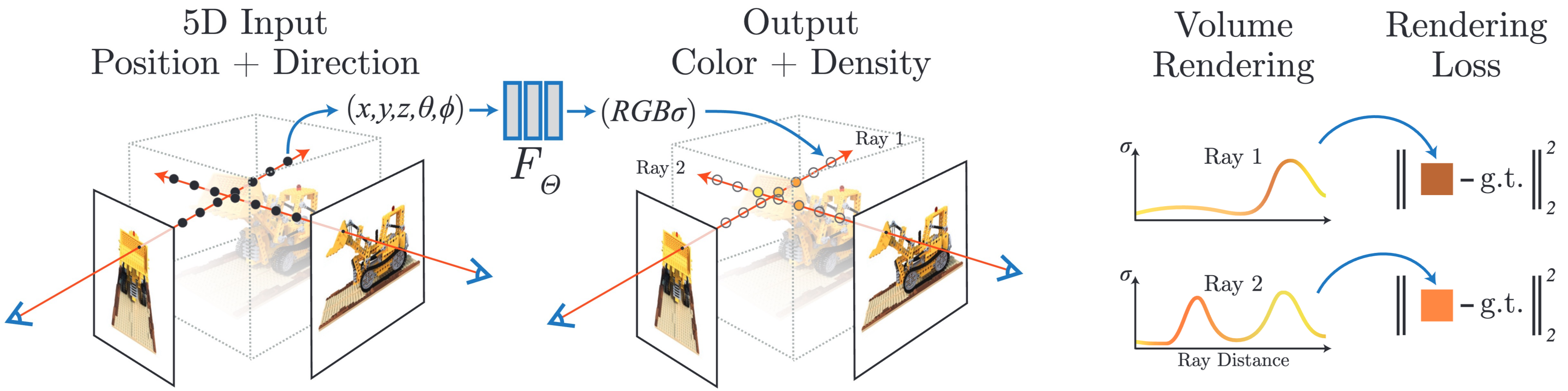
FLAME is a common 3DMM model used as a prior in most of 3D face avatar methods

$$\mathbf{S} = \bar{\mathbf{S}} + \delta \mathbf{B}_{\text{id}} + \gamma \mathbf{B}_{\text{expr}}$$

$$\mathbf{C} = \bar{\mathbf{C}} + \sigma \mathbf{D}_{\text{id}}$$

*Learning a model of facial shape and expression from 4D scans – Li *et al.*

NeRF*



$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{c}_i$$

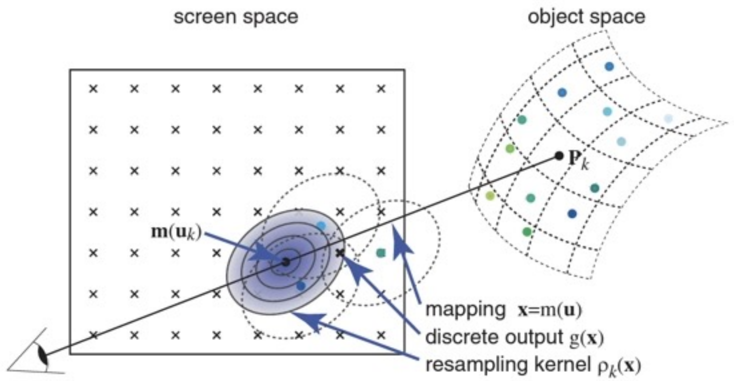
*Representing Scenes as Neural Radiance Fields for View Synthesis 3D– Mildenhall *et al.*

3DGS*

Primitives as ellipsis

Splatted Gaussians

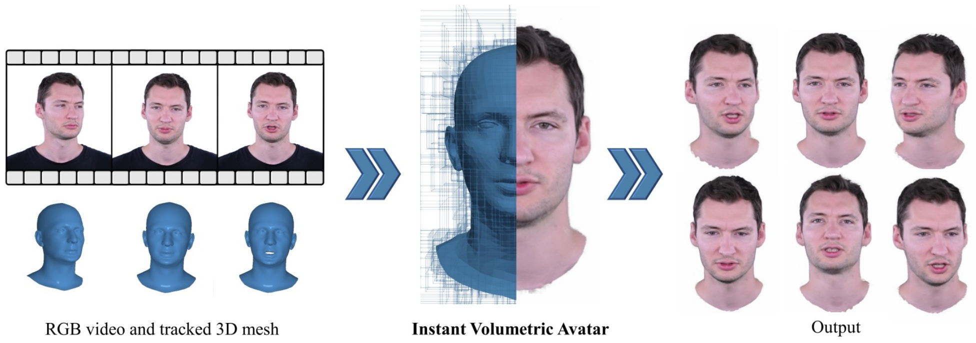
3D Gaussian Splatting by Kerbl et al.



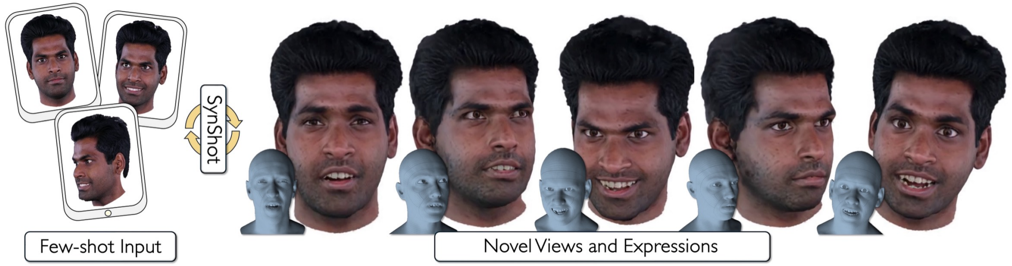
Surface Splatting by Zwicker et al.

*3D Gaussian Splatting for Real-Time Radiance Field Rendering – Kerbl *et al.*

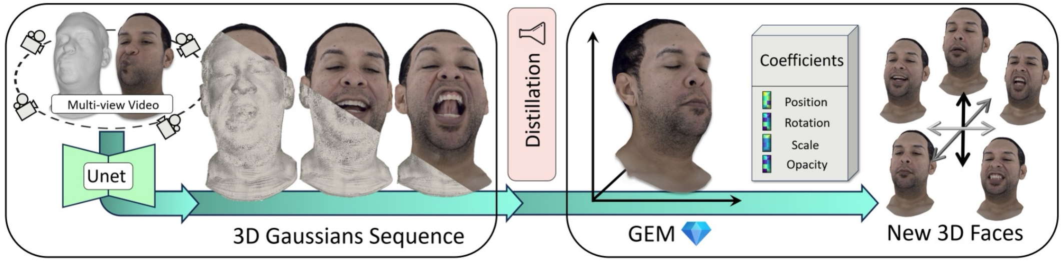
Presentation Outline



INSTA - Instant Volumetric Head Avatars
[Zielonka, Bolkart, Thies**]**
CVPR'23



SynShot - Synthetic Prior for Few-Shot Drivable Avatar Inversion
[Zielonka, Garbin, Lattas, Kopanas, Gotardo, Beeler, Thies, Bolkart**]**
CVPR'25



GEM - Gaussian Eigen Models for Human Heads
[Zielonka, Bolkart, Beeler, Thies**]**
CVPR'25

Motivation: Avatar from a monocular video

Motivation: Optimization Time

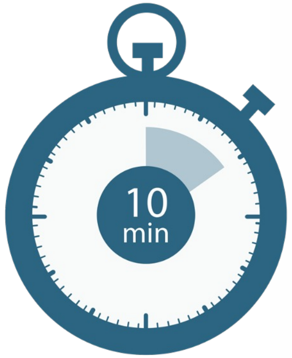
~ **2 days**
1 GPUs
NHA [Grassal *et al.*]

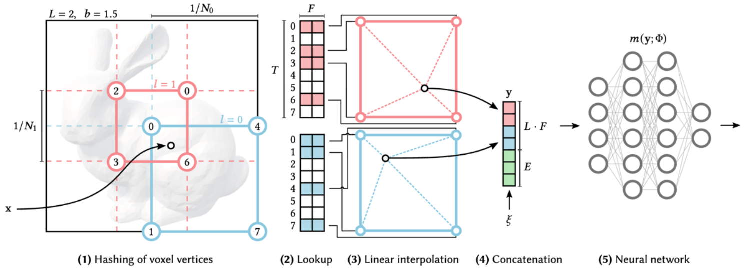
~ **3-4 days**
1 GPU
NeRFace [Gafni *et al.*]

~ **4-5 days**
1 GPU
IMAvatar [Zheng *et al.*]

Ground truth

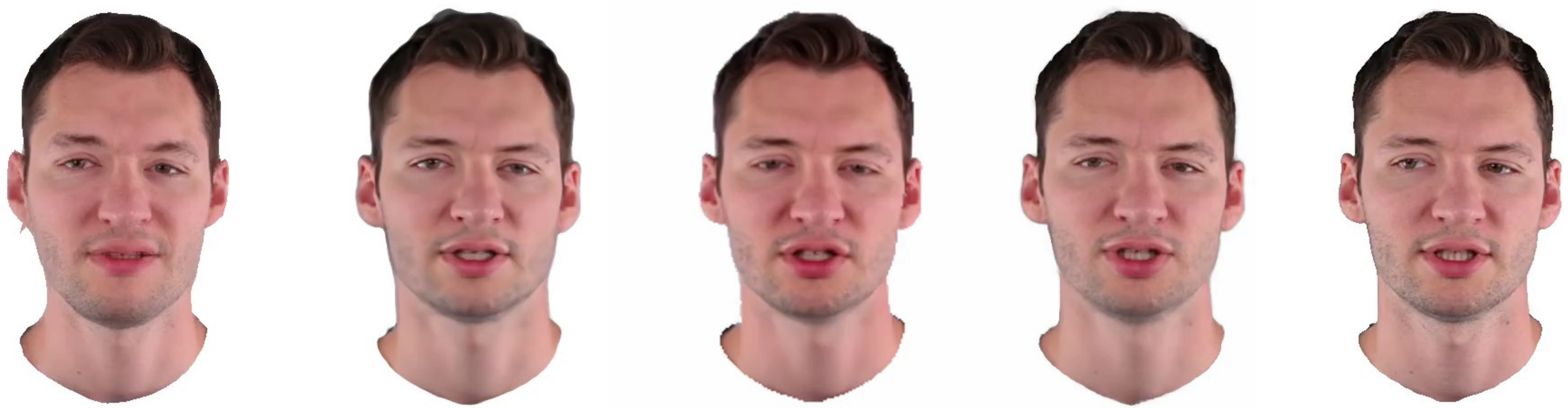
Solution: NeRF embedded on the surface of a mesh





Instant Neural Graphics Primitives with a Multiresolution Hash Encoding by Muller *et al.*

Solution: Order of magnitude time improvement



~ 2 days
1 GPUs
NHA [Grassal *et al.*]

~ 3-4 days
1 GPU
NeRFace [Gafni *et al.*]

~ 4-5 days
1 GPU
IMAvatar [Zheng *et al.*]

~ 10 minutes
1 GPU
Ours

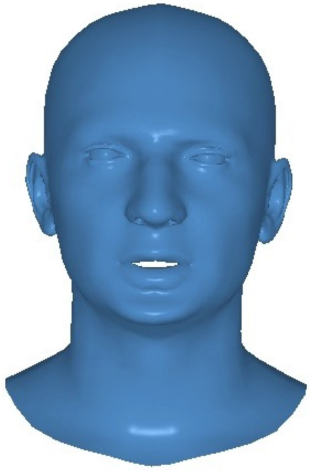
Ground truth

Method: Input (monocular video + tracked mesh)

RGB video

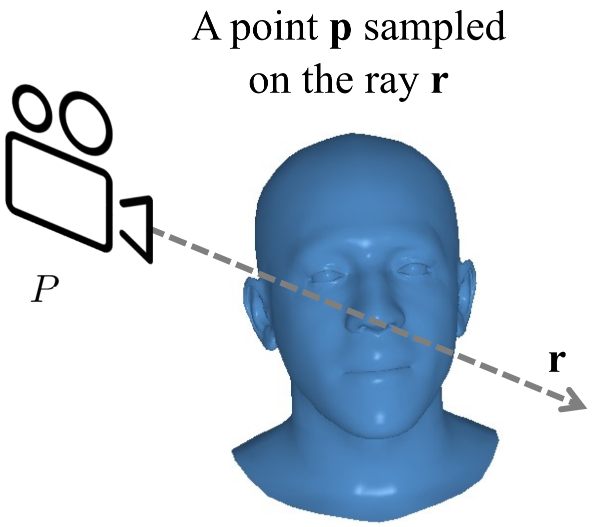
Landmarks

Statistical texture



Geometry

Method: Pipeline



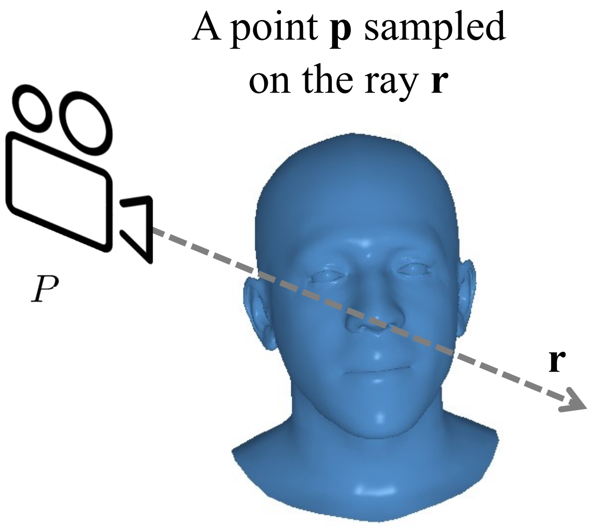
Tracked mesh M

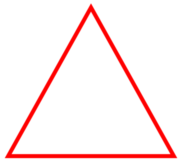
$\boldsymbol{E}$ – expressions

$\boldsymbol{P}$ – camera poses

$\triangle$ – nearest triangle to **p**

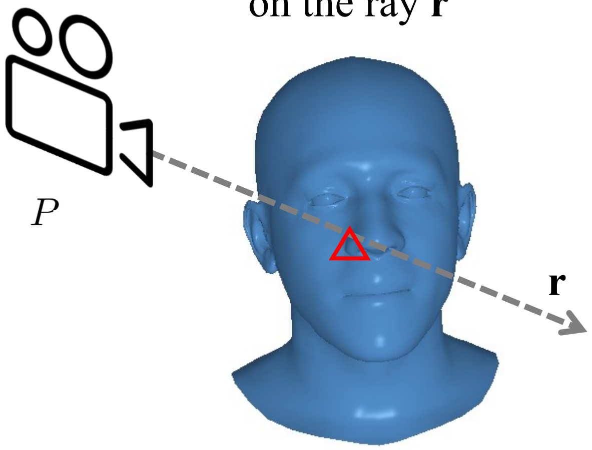
Method: Pipeline





Tracked mesh M
$\boldsymbol{E}$ – expressions
$\boldsymbol{P}$ – camera poses
△ – nearest triangle to **p**

Method: Pipeline



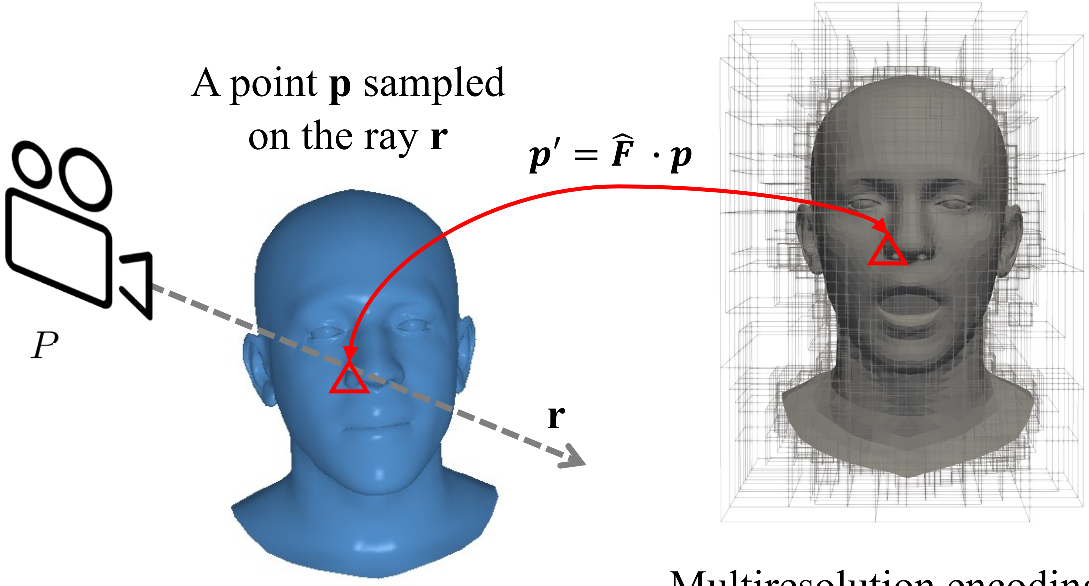
Tracked mesh M

$\boldsymbol{E}$ – expressions

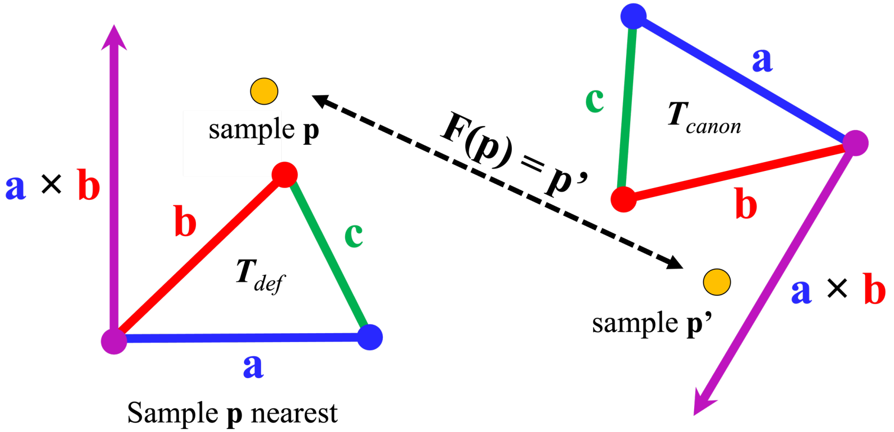
$\boldsymbol{P}$ – camera poses

$\triangle$ – nearest triangle to **p**

Method: Pipeline

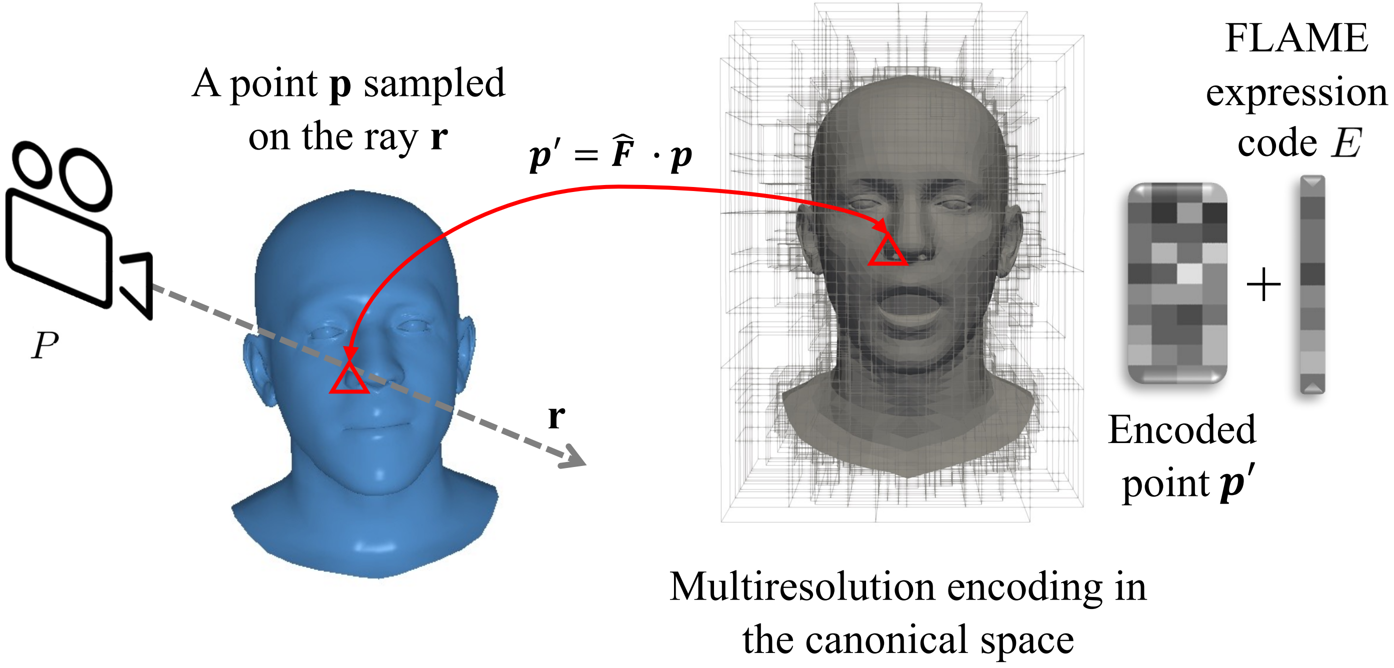


Multiresolution encoding in the canonical space



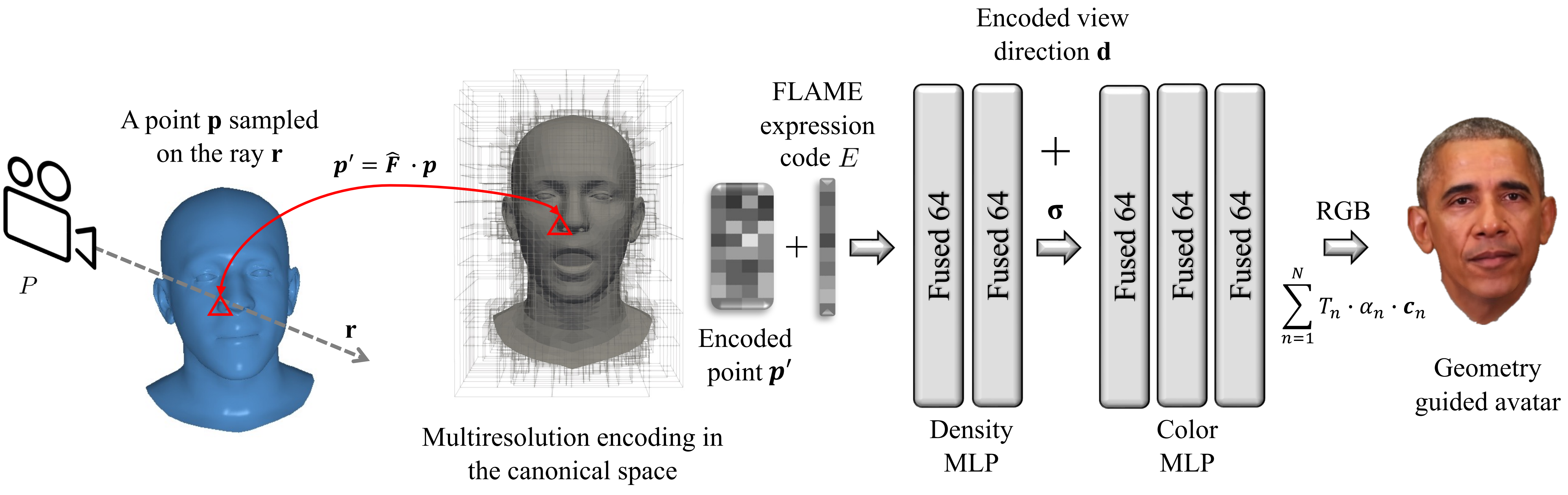
Tracked mesh M
$\boldsymbol{E}$ – expressions
$\boldsymbol{P}$ – camera poses
△ – nearest triangle to **p**

Method: Pipeline



Tracked mesh M
$\boldsymbol{E}$ – expressions
$\boldsymbol{P}$ – camera poses
△ – nearest triangle to **p**

Method: Pipeline



Tracked mesh M

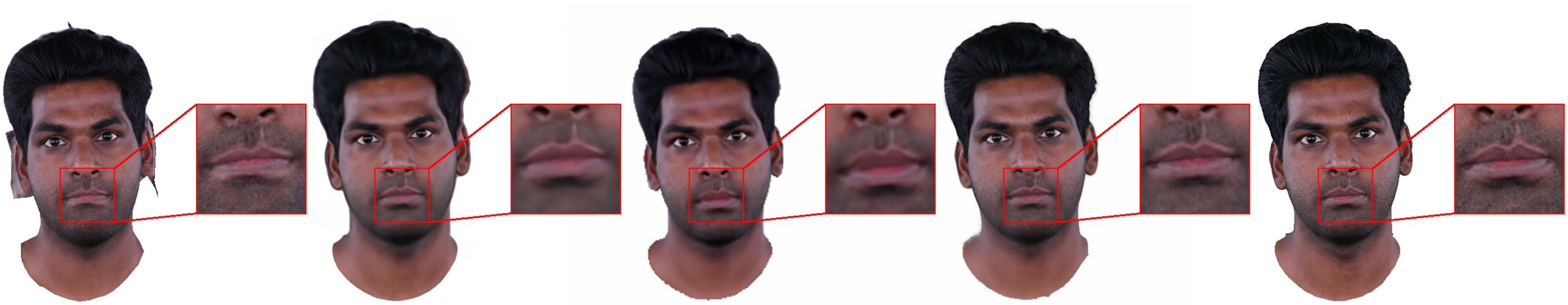
$\boldsymbol{E}$ – expressions

$\boldsymbol{P}$ – camera poses

△ – nearest triangle to **p**

Short Real-time Demo

Results: Qualitative Comparison



NHA
[Grassal et al.]

NeRFace
[Gafni et al.]

IMAvatar
[Zheng et al.]

Ours

Ground truth

Results: Quantitative Comparison

Method	L2 ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NHA [Grassal et al.]	0.0018	28.65	**0.96**	**0.03**
IMAvatar [Zheng et al.]	0.0014	29.10	0.95	0.06
NeRFace [Gafni et al.]	**0.0010**	**30.87**	**0.96**	0.05
Ours	**0.0010**	30.51	**0.96**	**0.03**

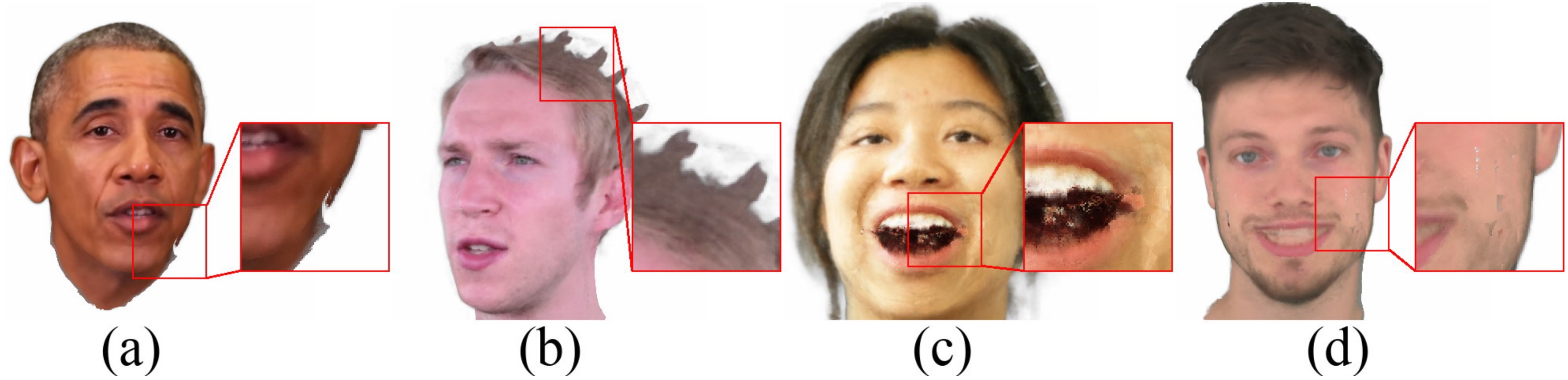
Results: Retargeting

INSTA works well for mostly small deformations but lacks a prior to properly generalize to other identities.

Source

Targets

Limitations

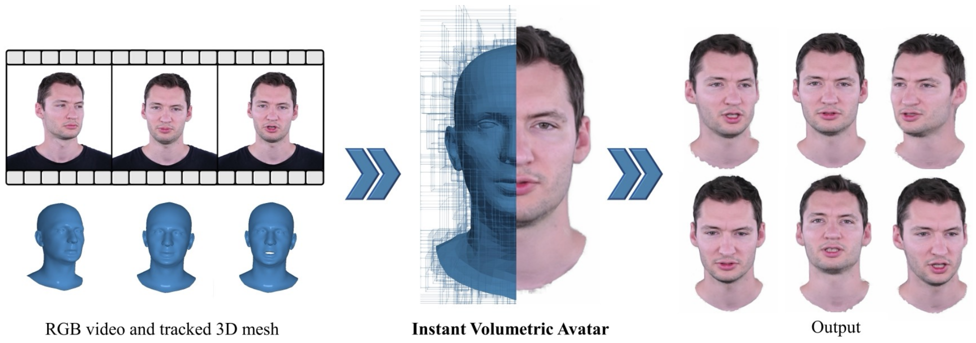


Failure cases: **(a)** and **(b)** exhibits outline artifacts at the chin and hair which stem from geometry misalignment of the tracker, **(c)** extreme expressions can cause artifacts in the mouth region, and **(d)** extrapolation of expressions can lead to artifacts.

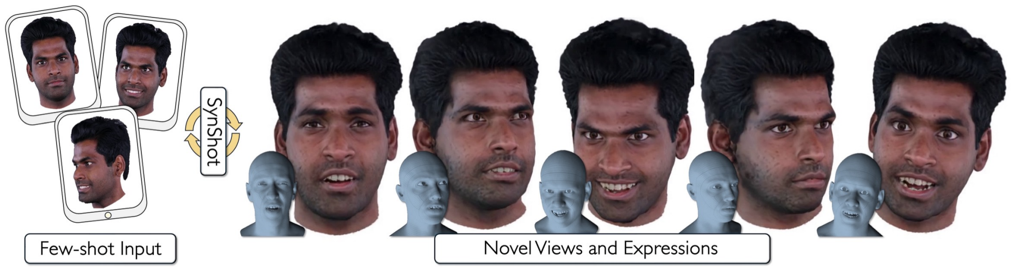
Take-home Messages of INSTA

1. Given a monocular RGB video as input, we optimize a controllable avatar in less than 10 minutes.
2. In this way, we can create a new avatar almost on the fly that reflects the current appearance instead of a prerecorded one.

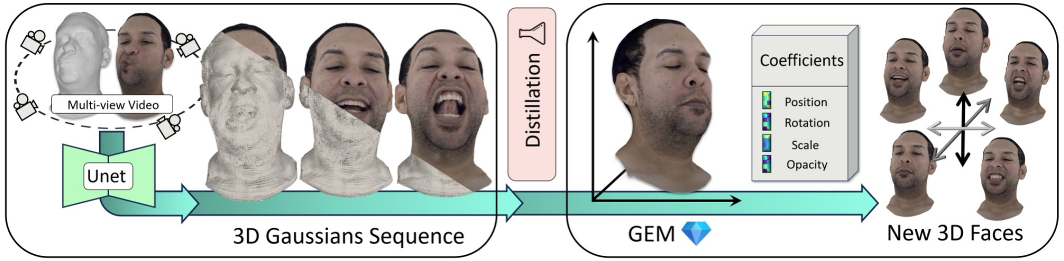
Presentation Outline



INSTA - Instant Volumetric Head Avatars
[Zielonka, Bolkart, Thies**]**
CVPR'23



SynShot - Synthetic Prior for Few-Shot Drivable Avatar Inversion
[Zielonka, Garbin, Lattas, Kopanas, Gotardo, Beeler, Thies, Bolkart**]**
CVPR'25



GEM - Gaussian Eigen Models for Human Heads
[Zielonka, Bolkart, Beeler, Thies**]**
CVPR'25

Motivation: Few-shot avatar inversion

Motivation: Monocular Methods

Source

INSTA[1]

Splatting Avatar[2]

Flash Avatar[3]

1) Zielonka *et al.* Instant Volumetric Head Avatars
2) Xiang *et al.* FlashAvatar: High-fidelity Head Avatar with Efficient Gaussian Embedding
3) Zhijing *et al.* SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting

Motivation: GAN-based Methods

Source

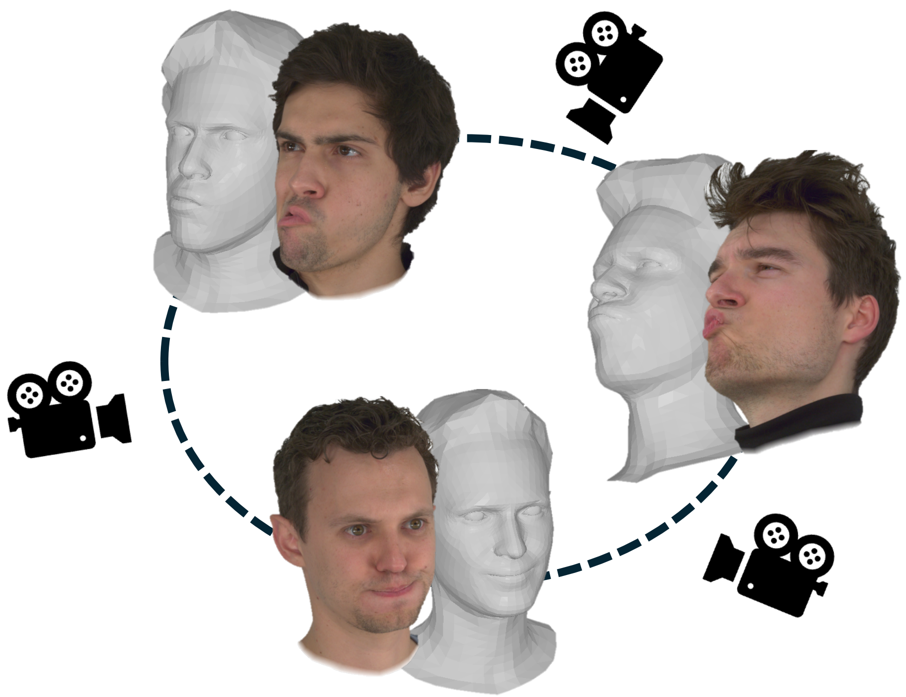
InvertAvatar[1]

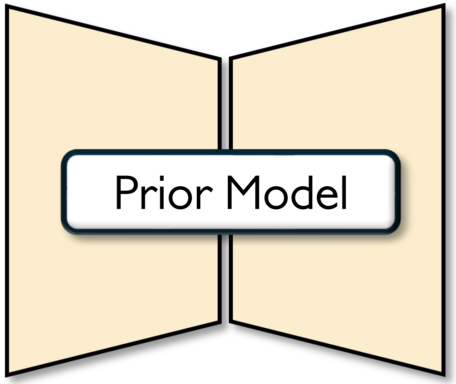
Portrait4D[2]

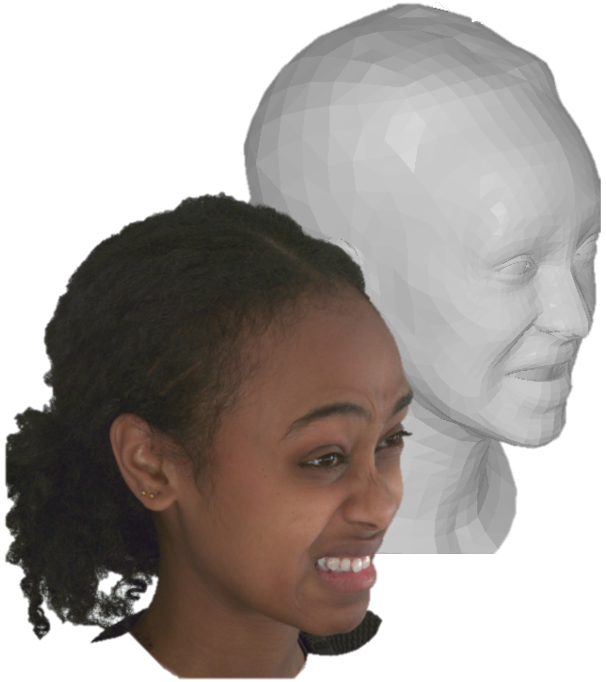
Next3D[3]

1) Zhao *et al.* InvertAvatar: Incremental GAN Inversion for Generalized Head Avatars
2) Deng *et al.* Portrait4D: Learning One-Shot 4D Head Avatar Synthesis using Synthetic Data
3) Sun *et al.* Next3D: Generative Neural Texture Rasterization for 3D-Aware Head Avatars

Solution: Train Prior Model







Use multi-view dataset with tracked meshes to build a prior model used for inversion and driving the avatars.

Solution: Real Datasets?

Nersemble[1]

Multiface[2]

FaceScape[3]

1) Kirschstein *et al.* NeRSemble: Multi-View Radiance Field Reconstruction of Human Heads
2) Wuu *et al.* Multiface: A Dataset for Neural Face Rendering
3) Zhu *et al.* FaceScape: 3D Facial Dataset and Benchmark for Single-View 3D Face Reconstruction

Solution: ~~Real~~ Datasets?

The **General Data Protection Regulation (GDPR)** is an EU law that protects individuals' personal data and privacy, enforced since May 25, 2018.

What does it mean for **Digital Humans research**:

1. Dataset derivatives must be frequently deleted e.g., each 30 days.
2. Trained models the same, periodically removed.

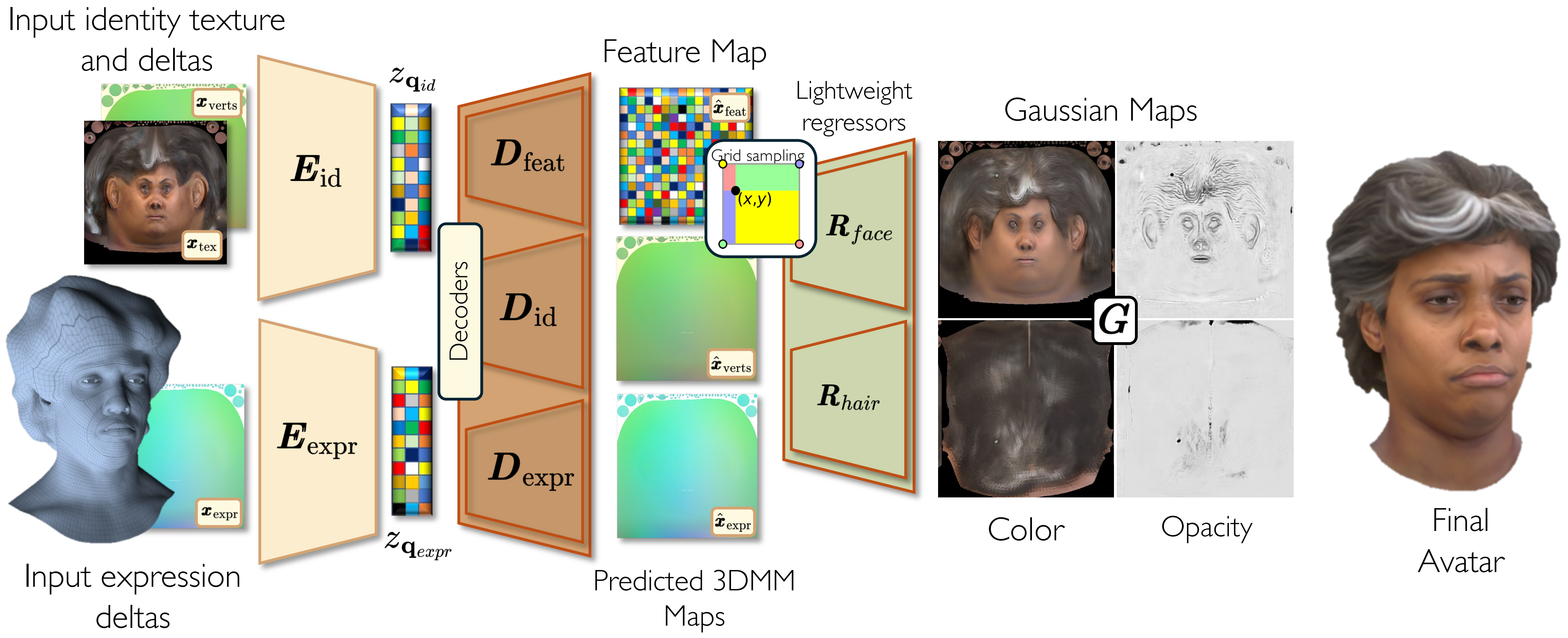
GDPR

Nersemble[1] Multiface[2] FaceScape[3]

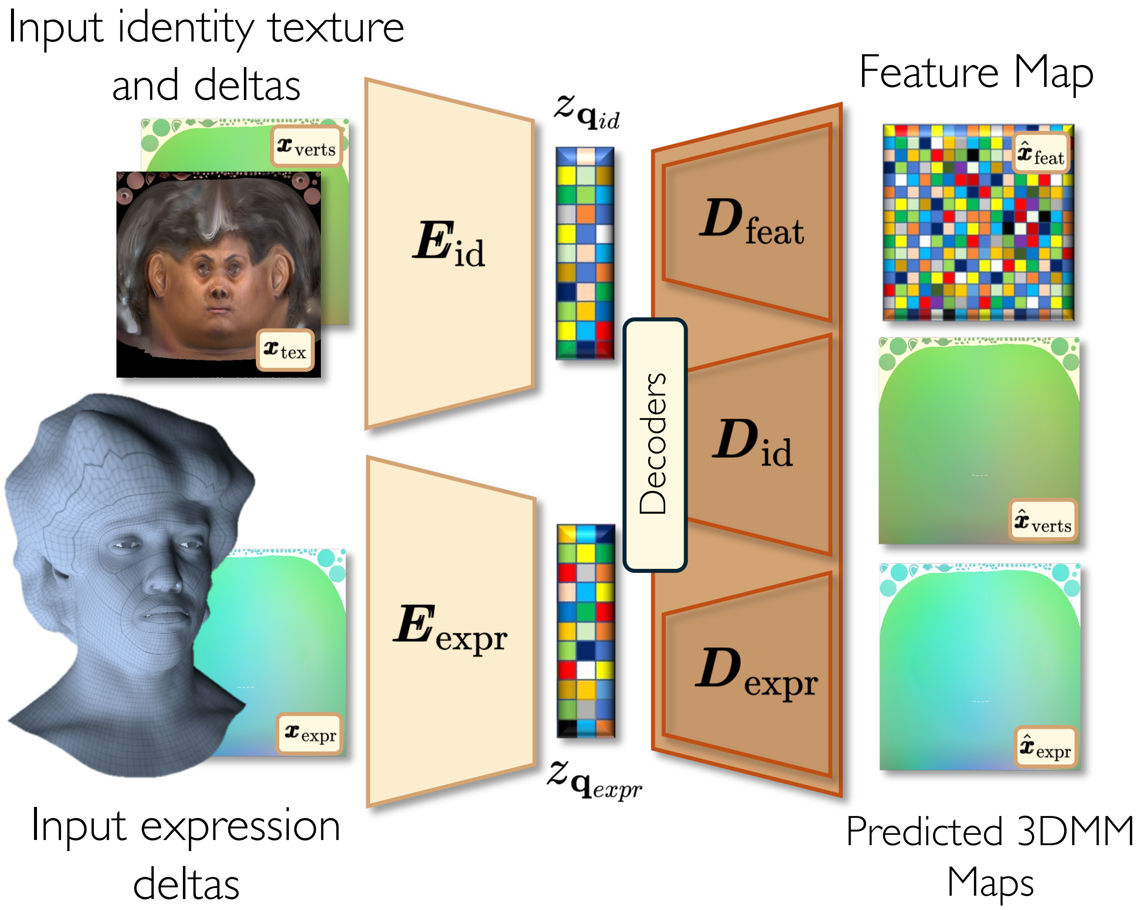
1) Kirschstein *et al.* NeRSemble: Multi-View Radiance Field Reconstruction of Human Heads
2) Wuu *et al.* Multiface: A Dataset for Neural Face Rendering
3) Zhu *et al.* FaceScape: 3D Facial Dataset and Benchmark for Single-View 3D Face Reconstruction

Solution: Synthetic Datasets ✓

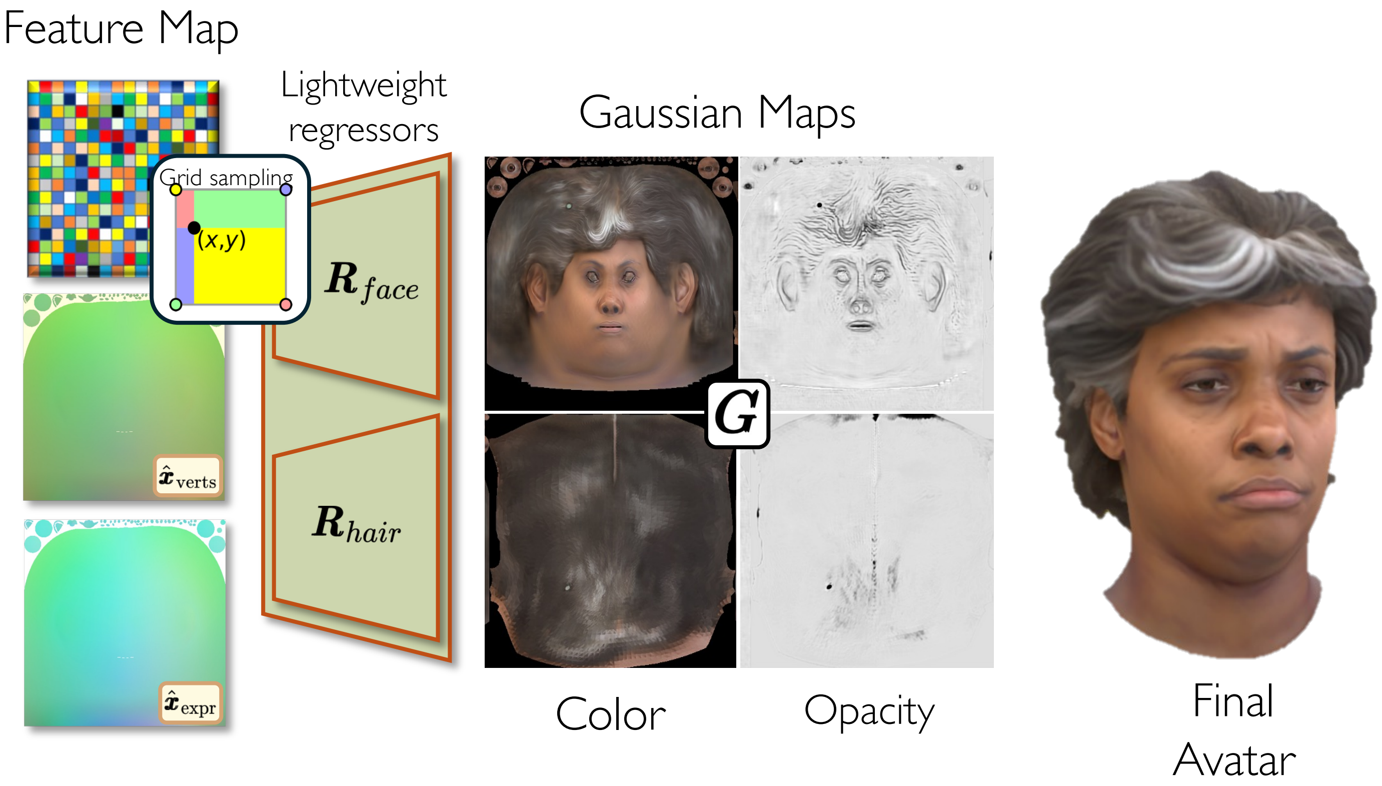
Method: Pipeline



Method: Encoder

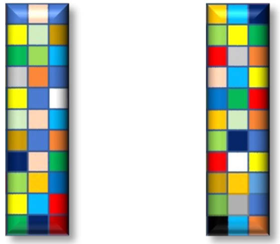


Method: Decoder



Method: Latent Space

$z_{\mathbf{q}_{id}}$ $z_{\mathbf{q}_{expr}}$



Interpolation

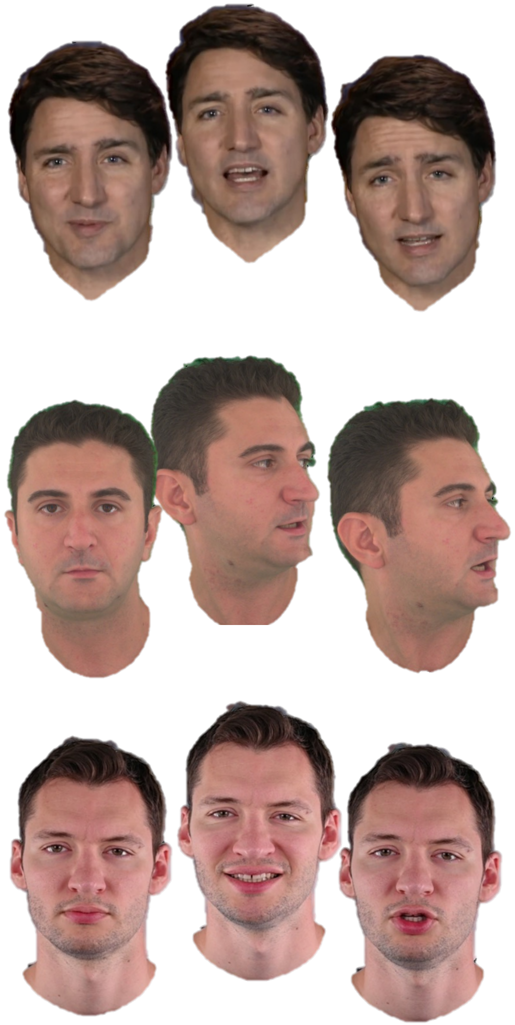
Method: Avatar Inversion

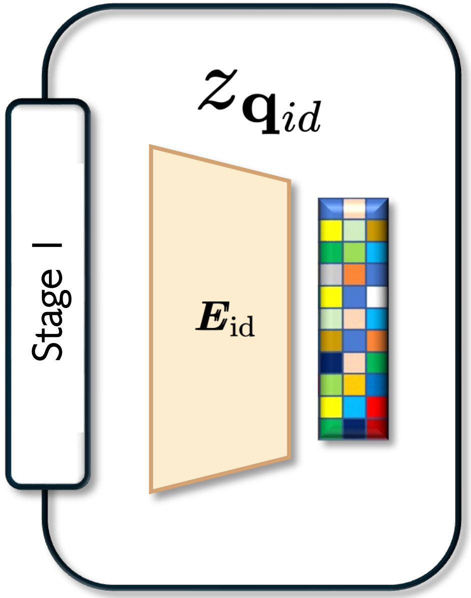
Few-shot Input

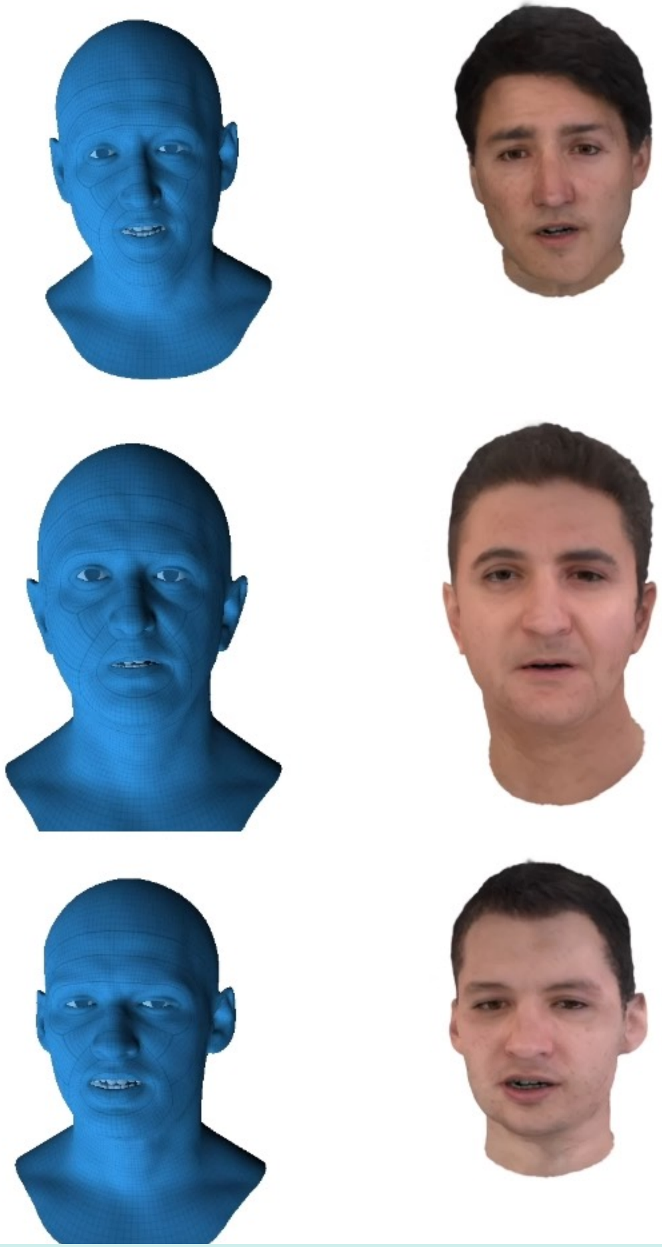
Inverted Avatar

Inversion – Stage 1

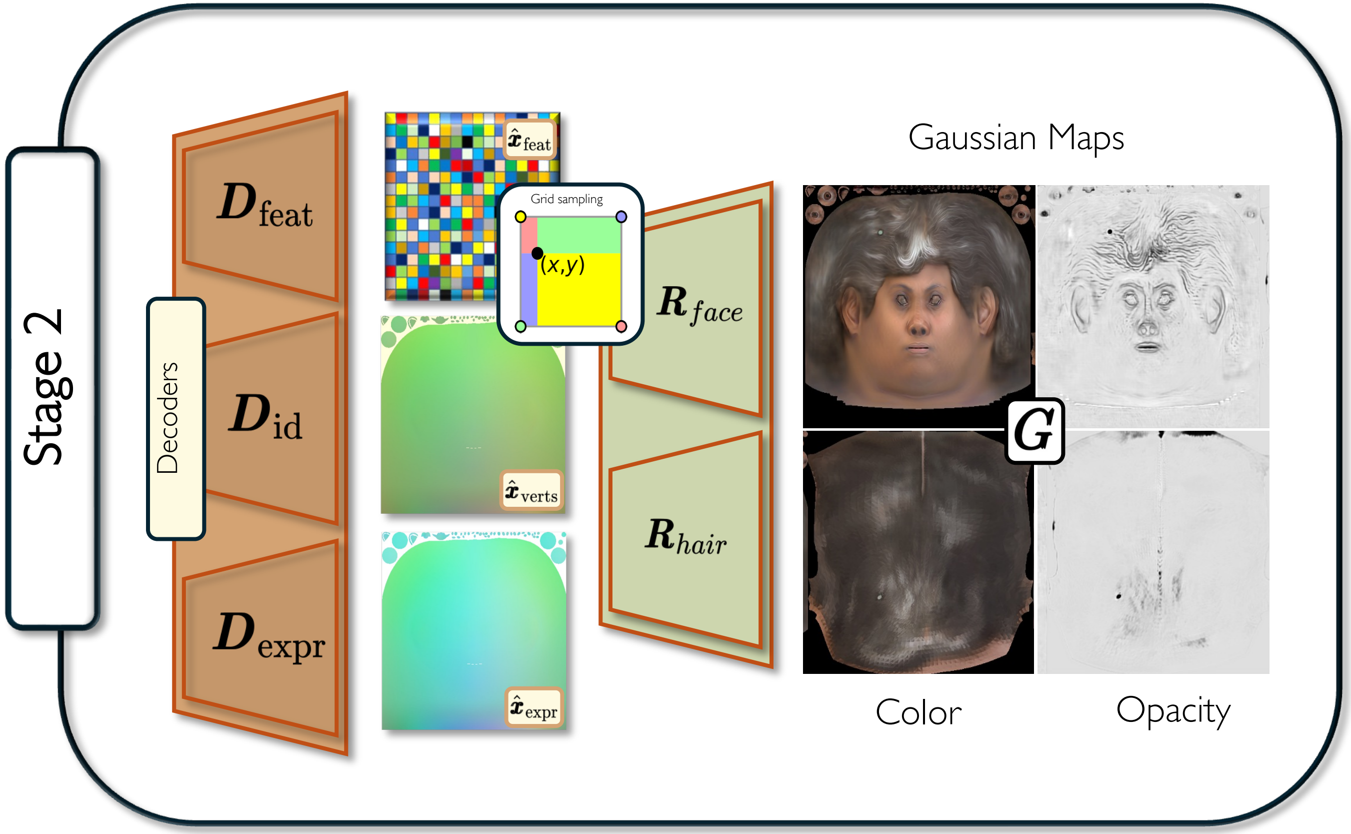
Input

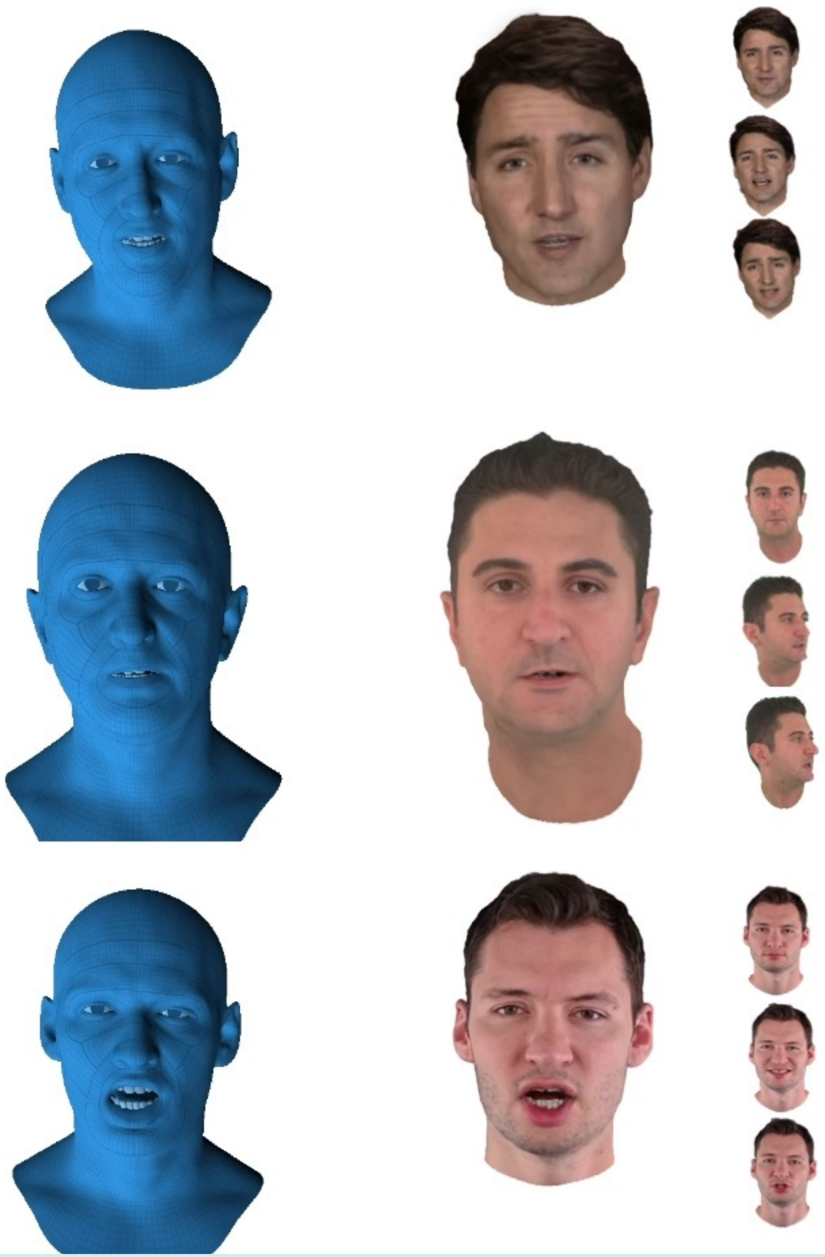




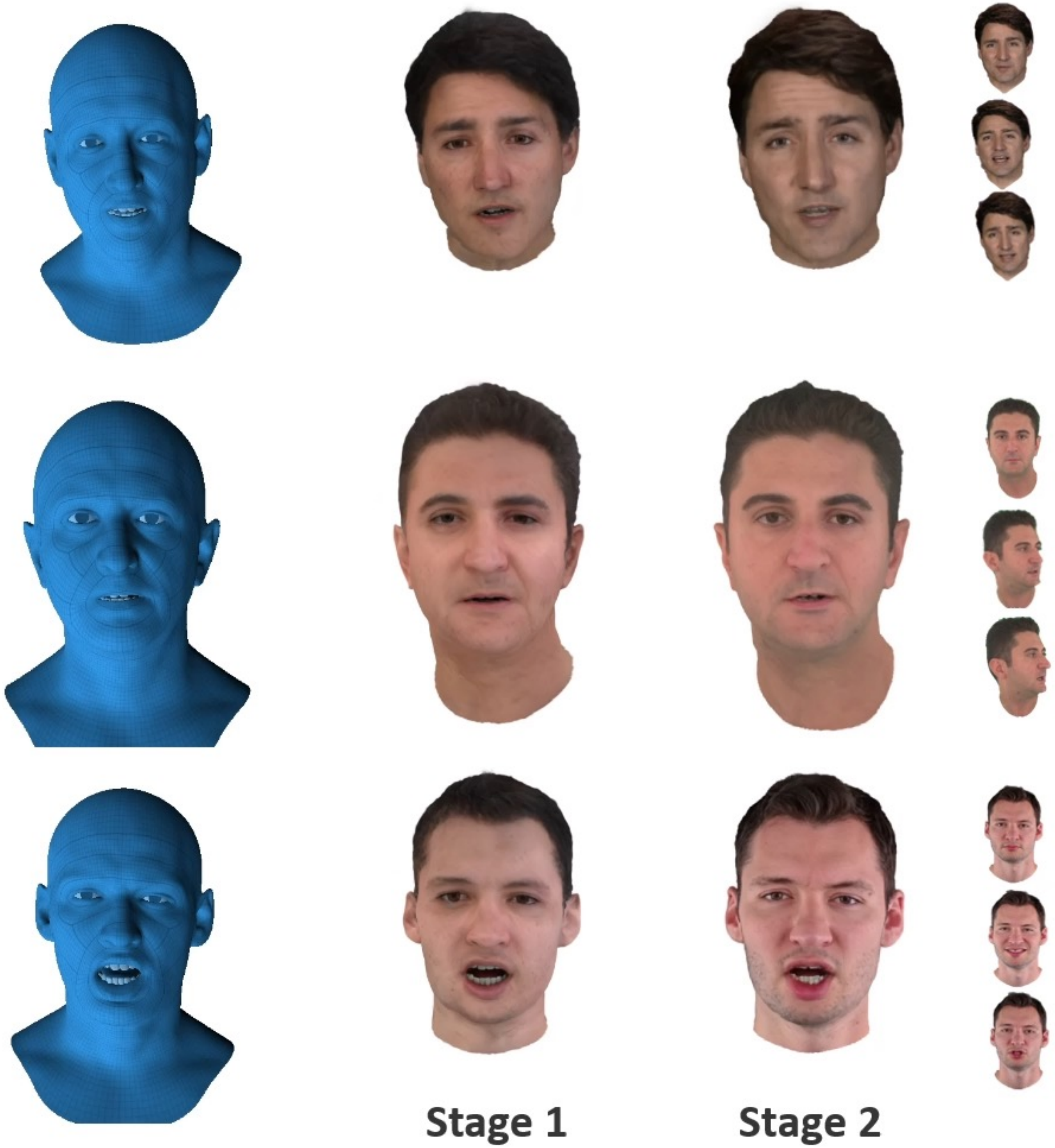


Inversion – Stage 2





Inversion – Stage 1-2



SynShot

Results: Personalized Baselines

Source Ours INSTA[1] Splatting Avatar[2] Flash Avatar[2]

1) Zielonka *et al.* Instant Volumetric Head Avatars
2) Xiang *et al.* FlashAvatar: High-fidelity Head Avatar with Efficient Gaussian Embedding
3) Zhijing *et al.* SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting

Results: GAN-based Baselines

Source | Ours | InvertAvatar[1] | Portrait4D[2] | Next3D[3]

1) Zhao *et al*. InvertAvatar: Incremental GAN Inversion for Generalized Head Avatars
2) Deng *et al*. Portrait4D: Learning One-Shot 4D Head Avatar Synthesis using Synthetic Data
3) Sun *et al*. Next3D: Generative Neural Texture Rasterization for 3D-Aware Head Avatars

Results: Three-shot Inversions

Results: Three-shot Inversions

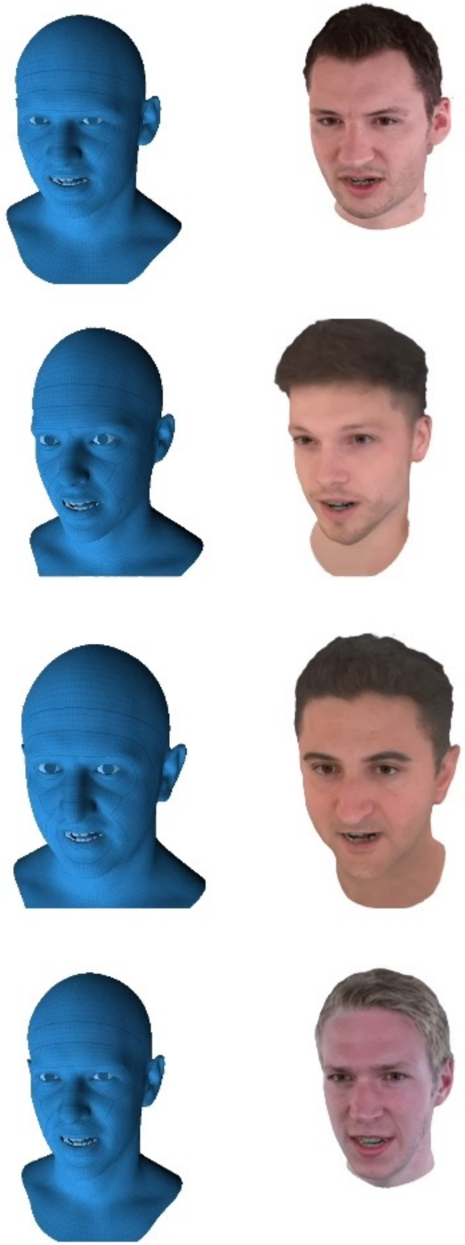
Limitations

Missing hair

No glasses in the dataset

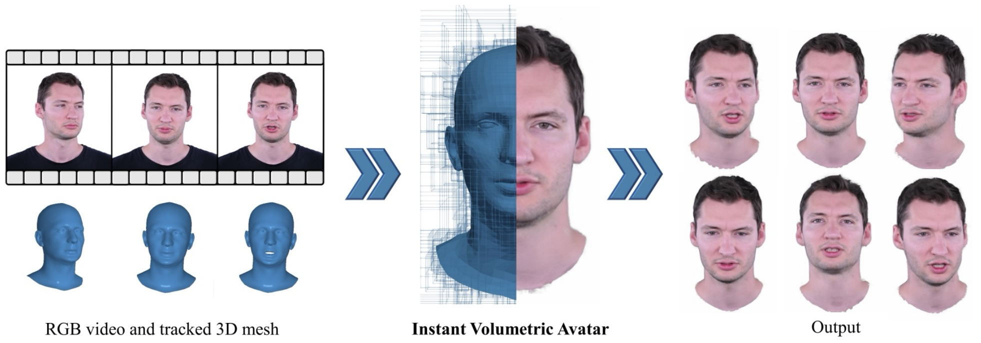
Wrong illumination

Take-home Messages of SynShot

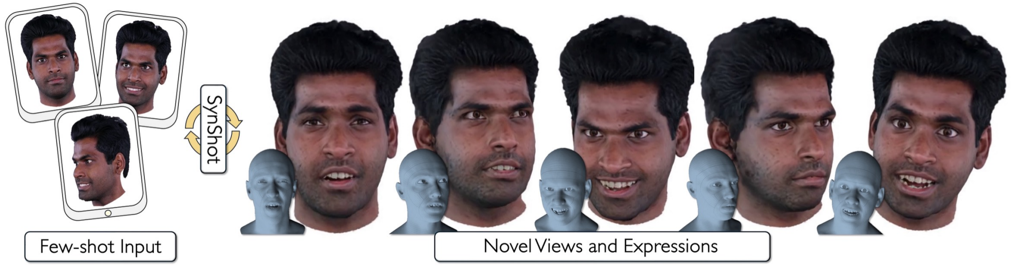


1. **SynShot** builds a multi-view prior using only ***synthetic*** data.
2. Enables cross-reenactment and outperforms monocular personalized methods like **INSTA**.
3. Pivotal fine-tuning bridges the real-synthetic gap, enabling drivable 4D avatars from synthetic-only priors.

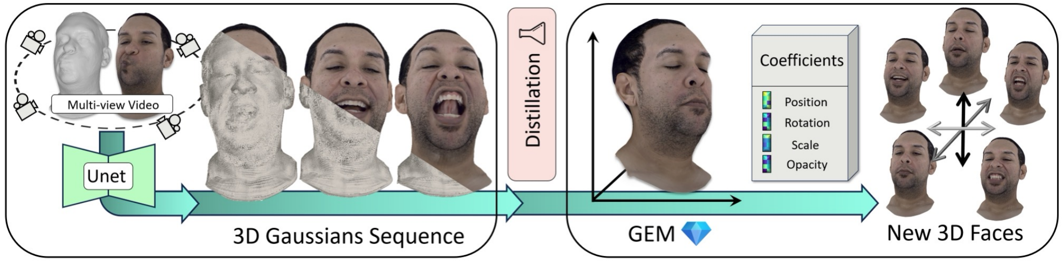
Presentation Outline



INSTA - Instant Volumetric Head Avatars
[Zielonka, Bolkart, Thies**]**
CVPR'23



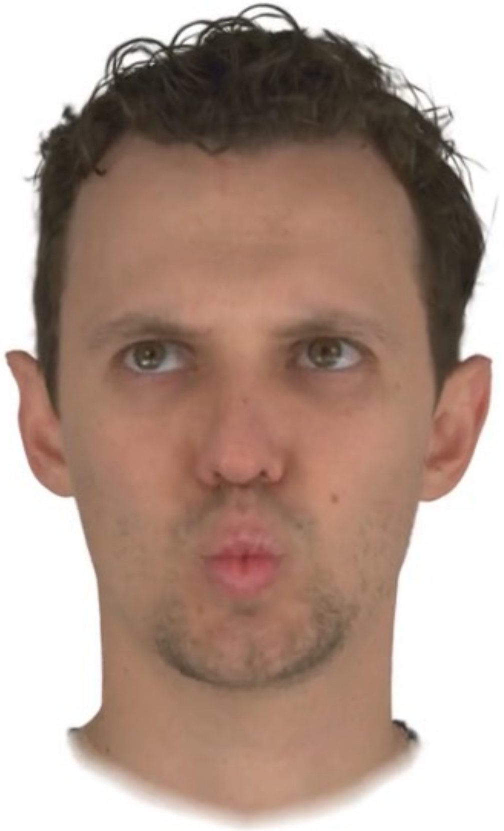
SynShot - Synthetic Prior for Few-Shot Drivable Avatar Inversion
[Zielonka, Garbin, Lattas, Kopanas, Gotardo, Beeler, Thies, Bolkart**]**
CVPR'25



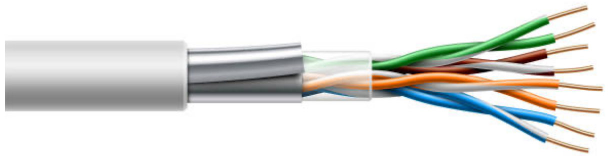
GEM - Gaussian Eigen Models for Human Heads
[Zielonka, Bolkart, Beeler, Thies**]**
CVPR'25

Motivation: High-Quality Avatars on VR-Glasses

Place A



GEM💎Avatar

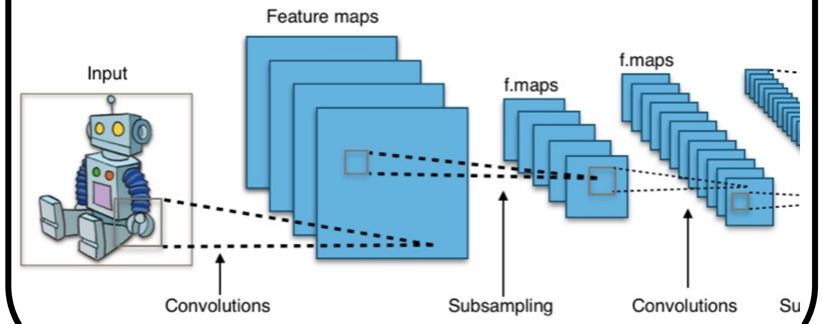


Place B

GEM💎Avatar

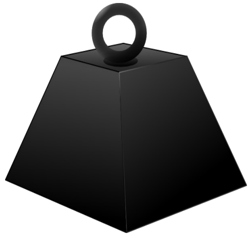
Motivation: There is no Free Lunch

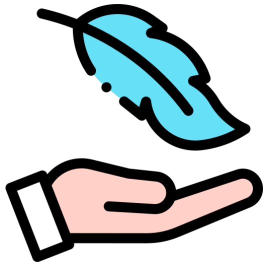
CNN Networks



3DGS

Gaussian Avatars [1]

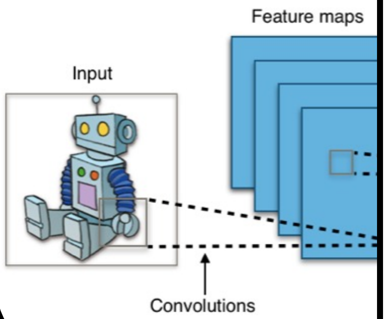


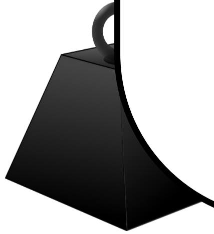


1) Qian et al., *Photorealistic Head Avatars with Rigged 3D Gaussians*, CVPR 2024

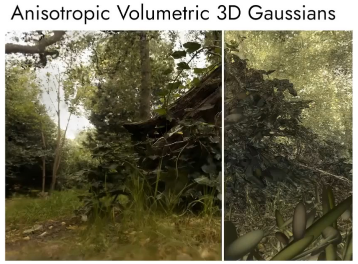
Motivation: There is no Free Lunch

CNN N



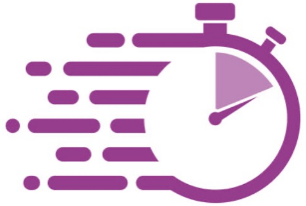


GS



Tradeoffs

Speed/weight (CNN models)



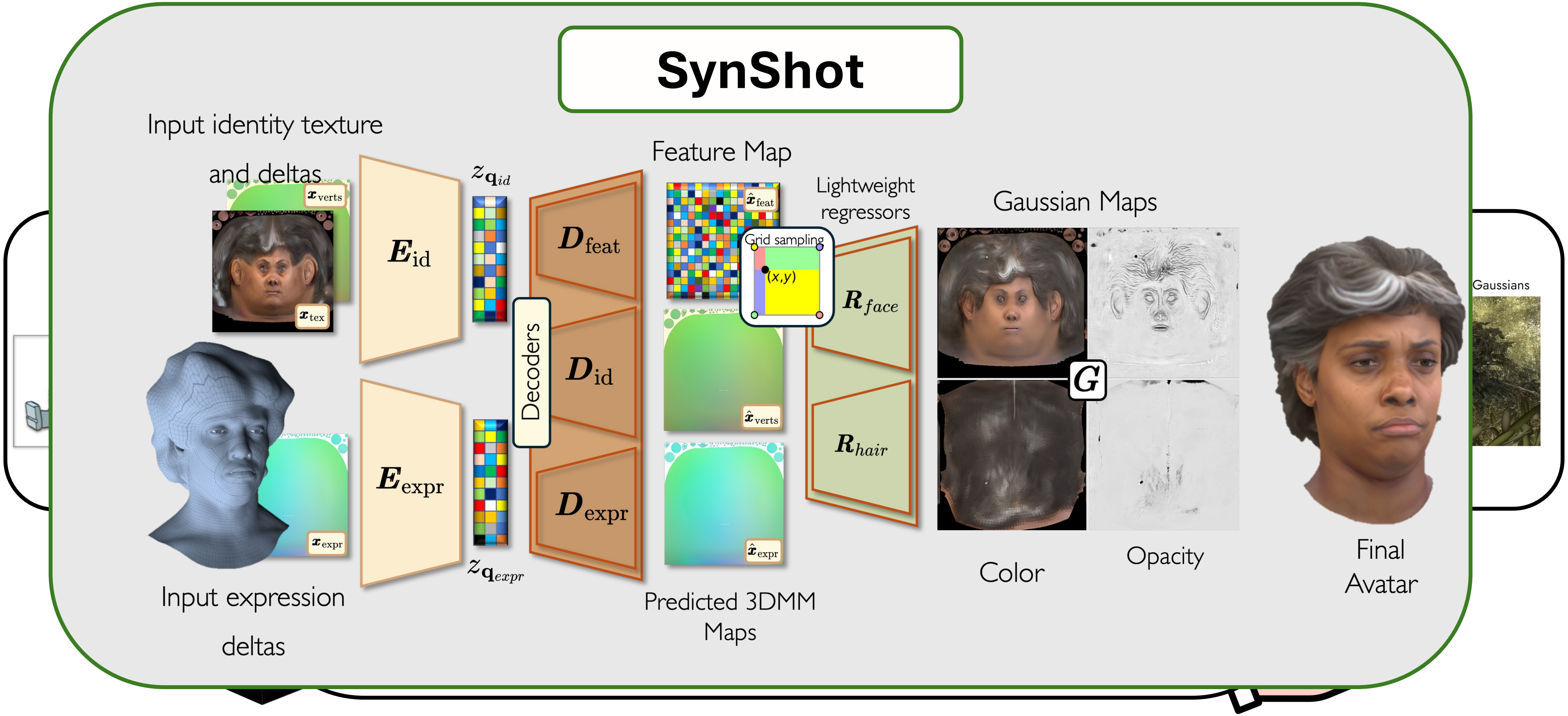
Quality (Lightweight models)

Ground truth

Gaussian Avatars [1]

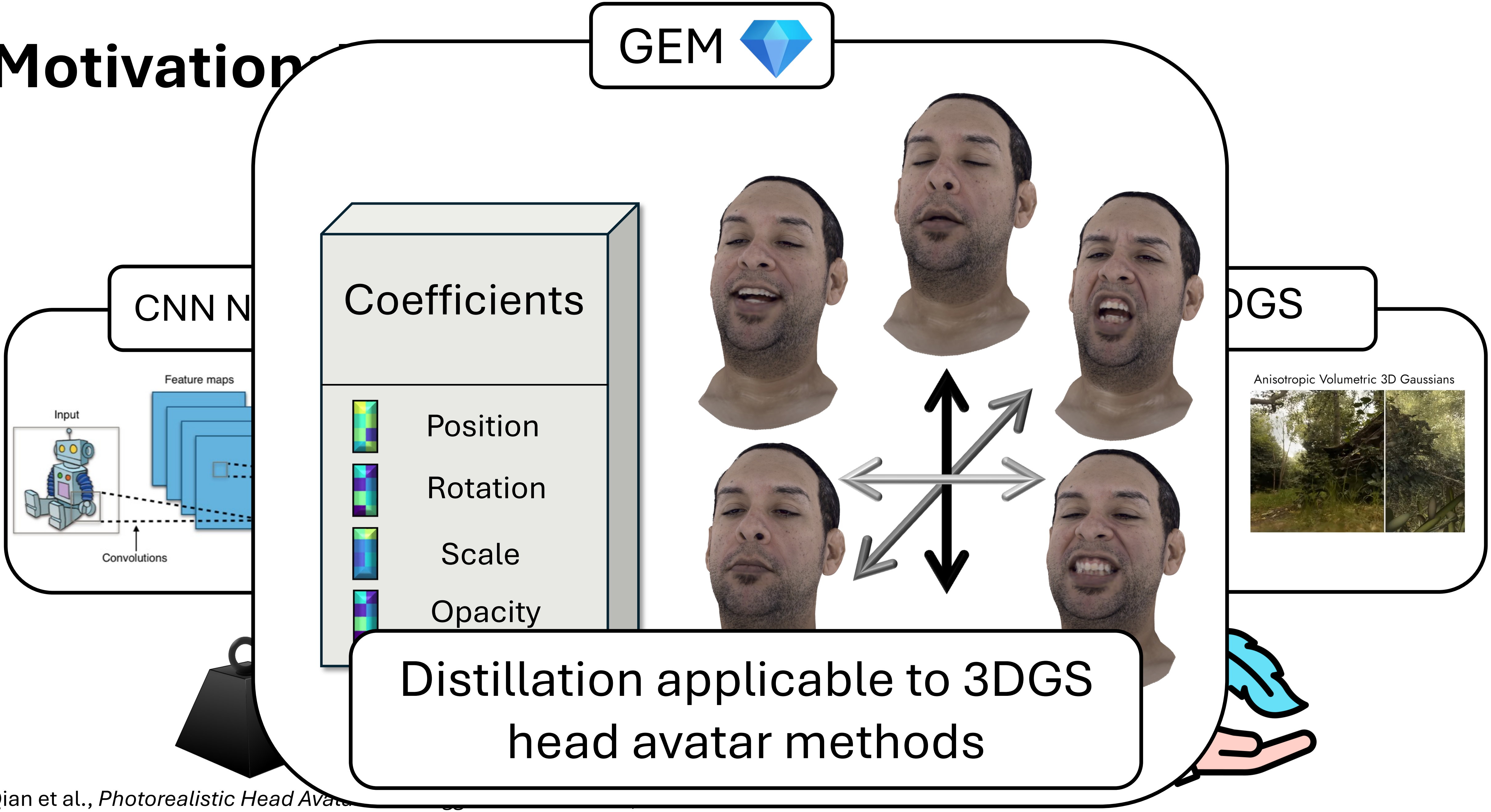
1) Qian et al., *Photorealistic Head Avatars with Rigged 3D Gaussians*, CVPR 2024

Motivation: There is no Free Lunch



1) Qian et al., *Photorealistic Head Avatars with Rigged 3D Gaussians*, CVPR 2024

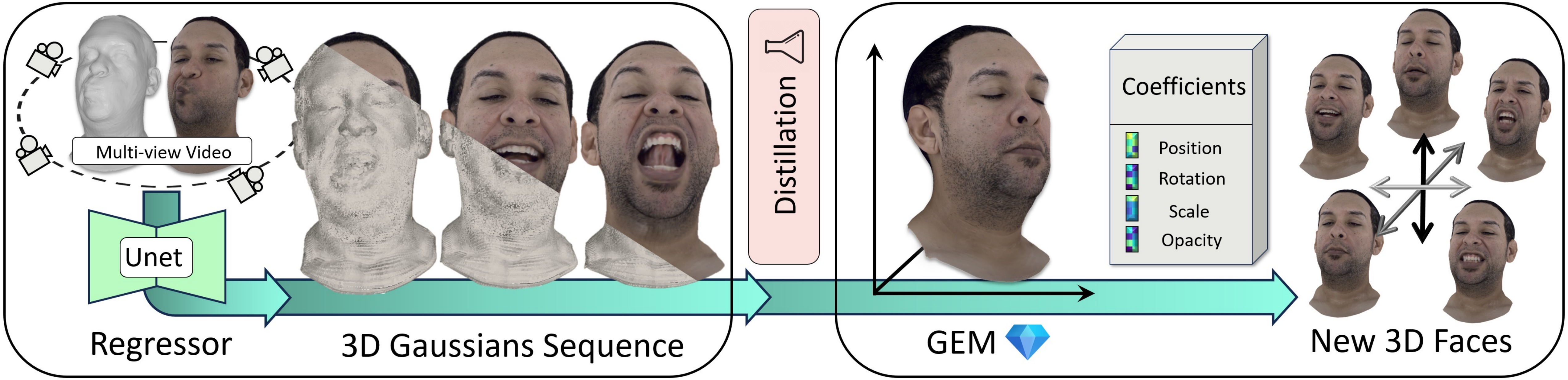
Motivation



Distillation applicable to 3DGS head avatar methods

1) Qian et al., *Photorealistic Head Ava*

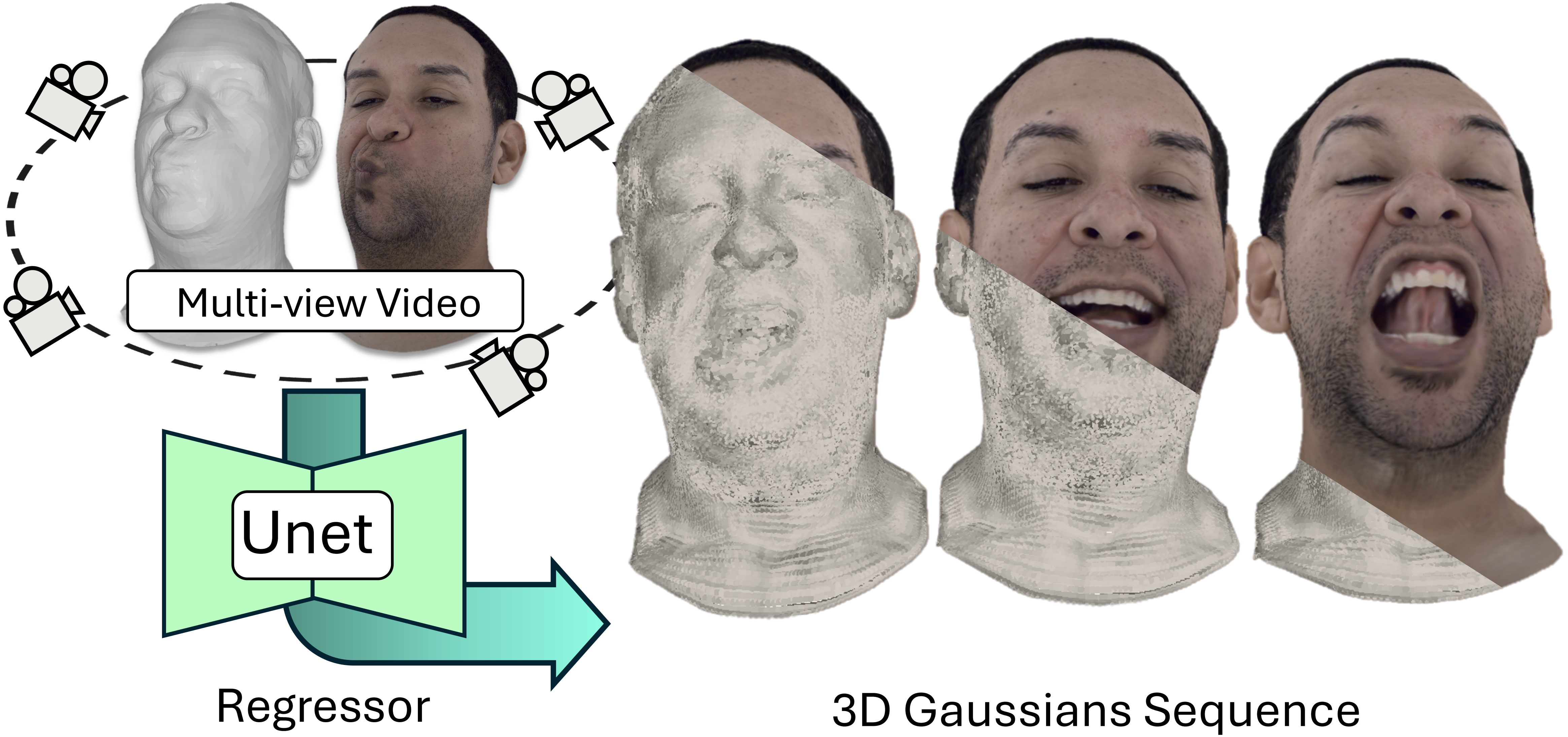
Method: Two Stage Approach



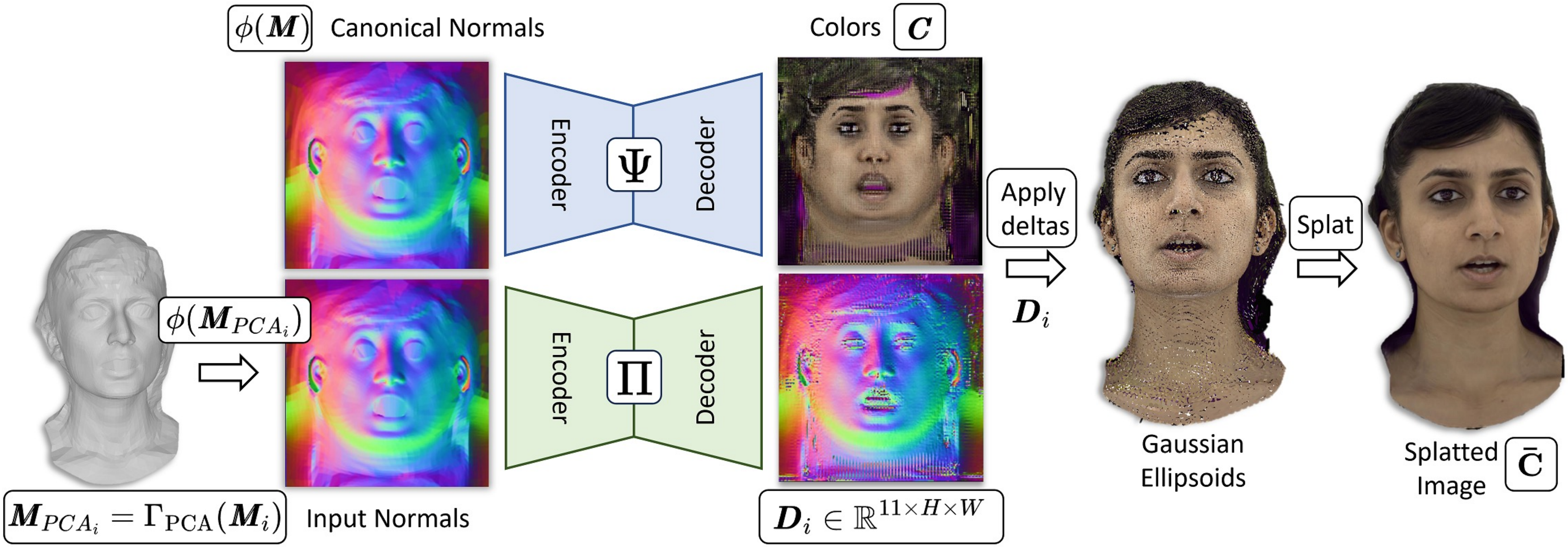
Stage 1

Stage 2

Method: Stage One (Generator)

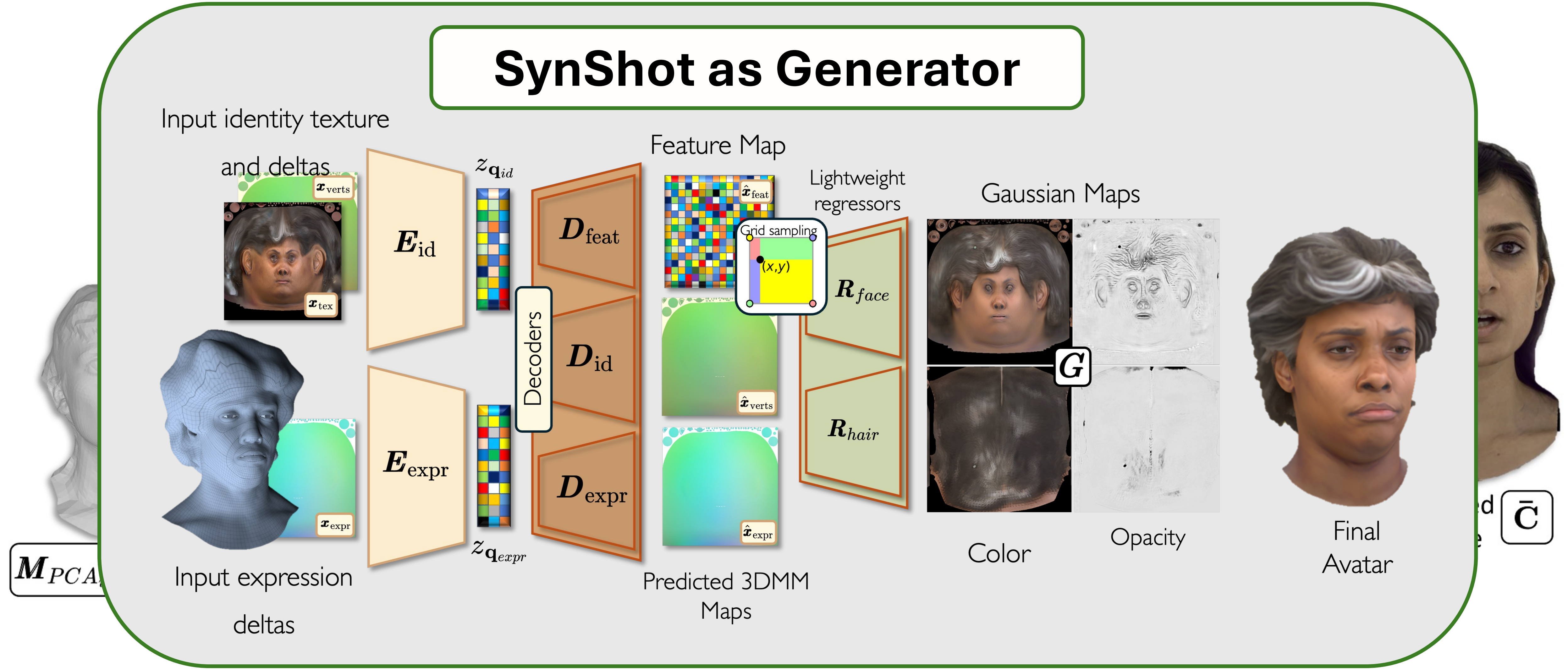


Method: Stage One (Sequence of Gaussians)

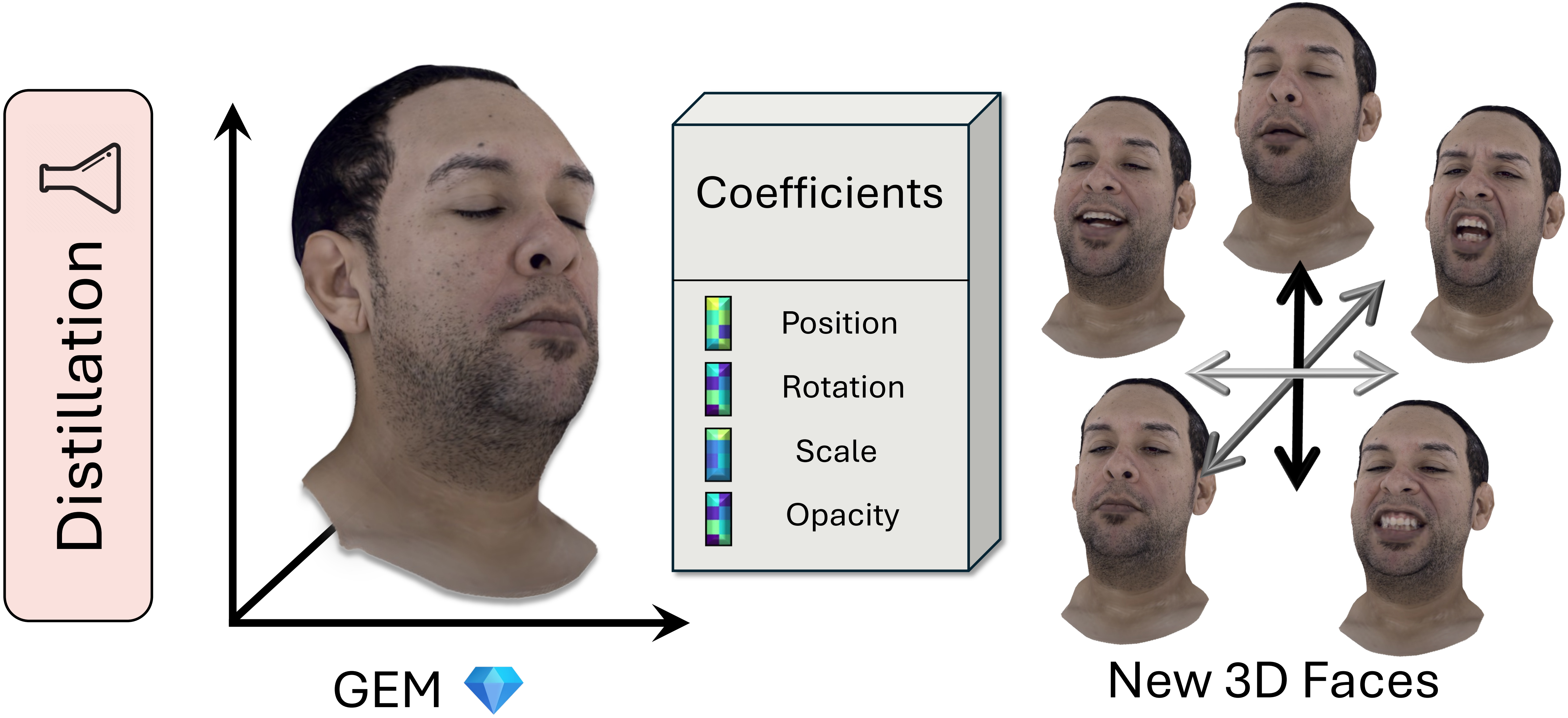


Animatable Gaussians [Li et al.]

Method: Stage One (Sequence of Gaussians)



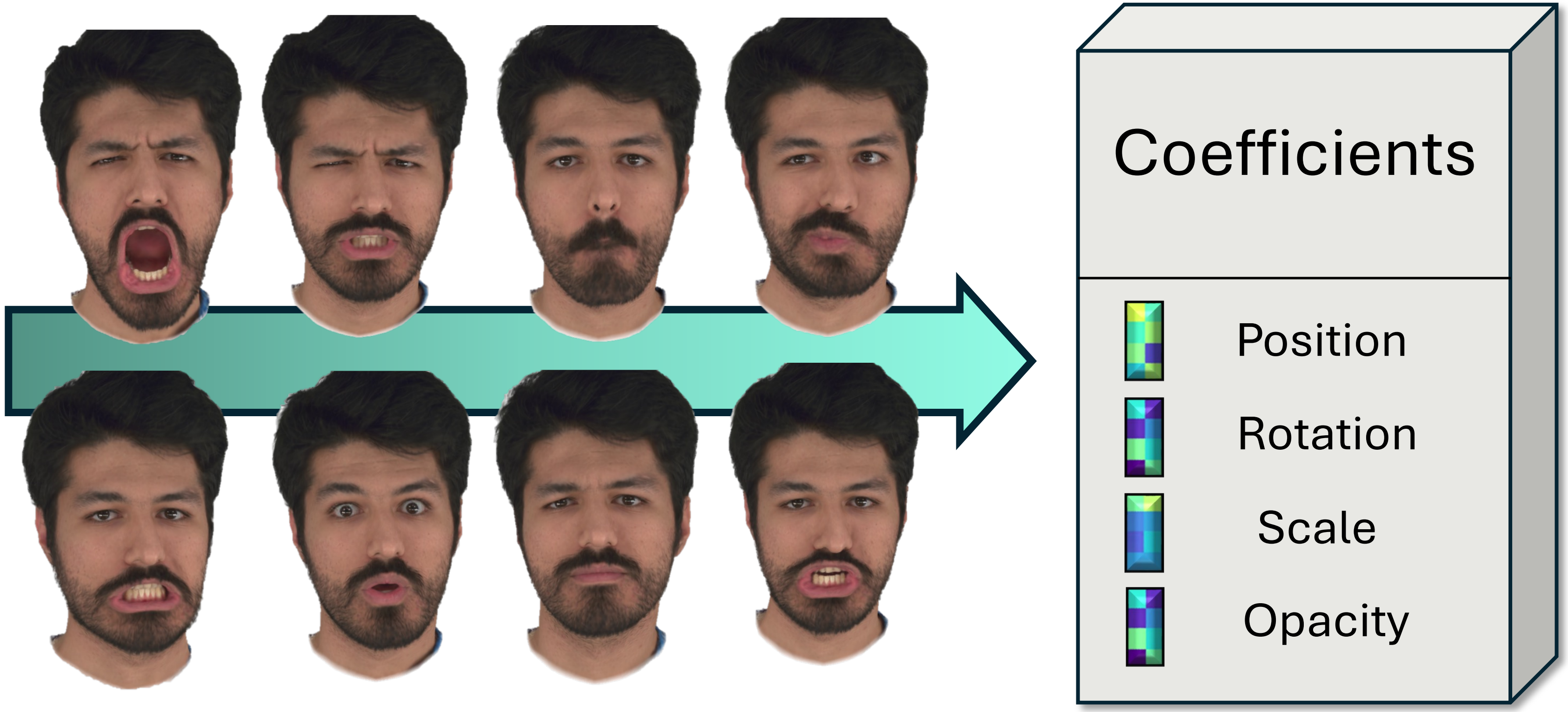
Method: Stage Two (Distillation)



Method: Stage Tw

Distilling

Distillation



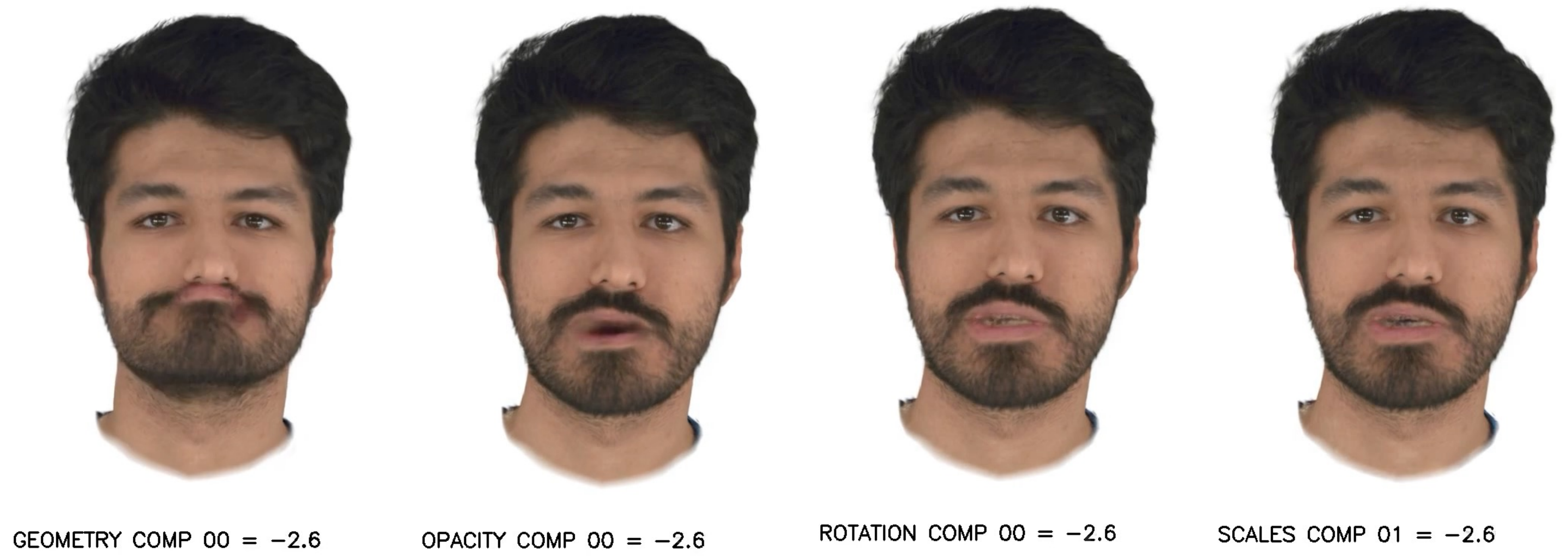
Dataset of regressed
Gaussians primitives

Faces

Method: Stage Tw

Ensamble of eigen basis

Distillation



Position Opacity Rotation Scale

Space traversal [-3σ, 3σ]

Faces

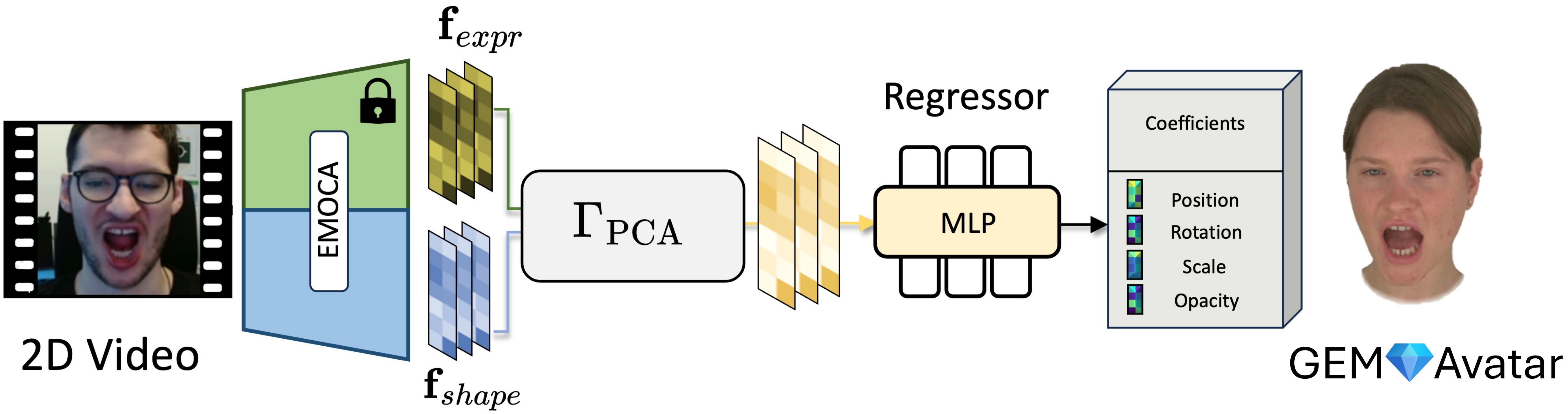
Results: Novel Expressions

AG – Animatable Gaussians [Li et al.]
GA – Gaussian Avatars [Qian et al.]
INSTA – [Zielonka et al.]

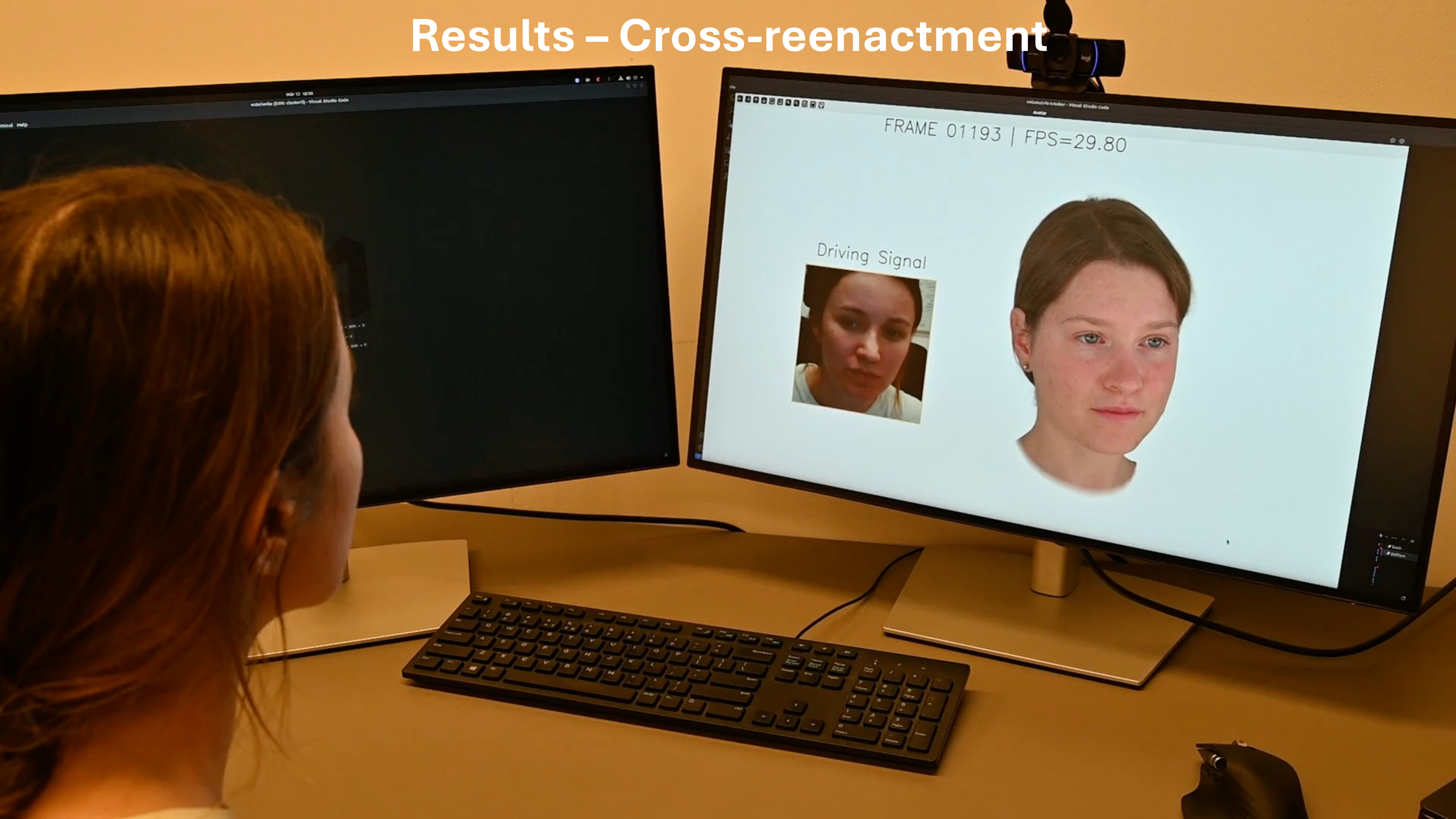
🔥 Needs a 3DMM like FLAME

Ground Truth | Ours 💎 GEM | Ours Net 🔥 | AG 🔥 | GA 🔥 | INSTA 🔥

Results: Image-base Cross-reenactment

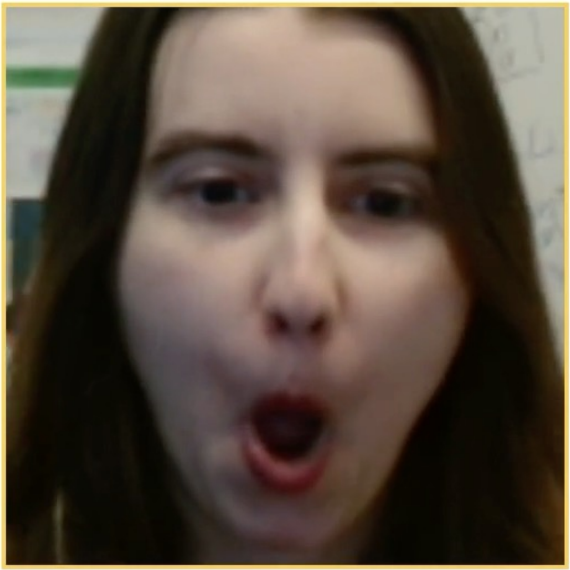


Results – Cross-reenactment



FRAME 03669 | FPS=29.60

Driving Signal



FRAME 00279 | FPS=30.15

Driving Signal

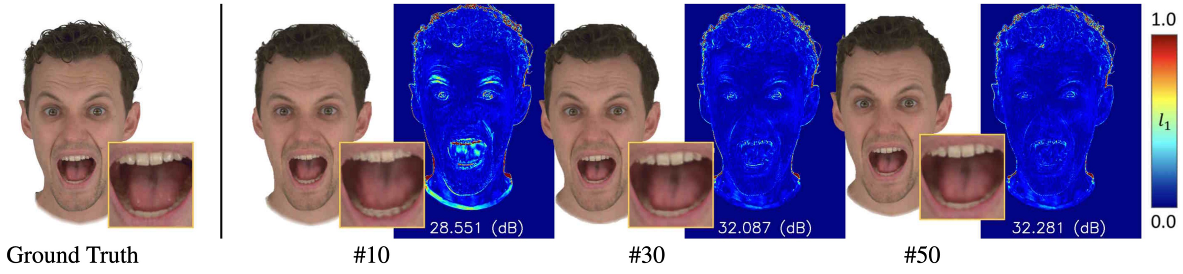
Limitations: PCA's Global Extend

Source

GEM

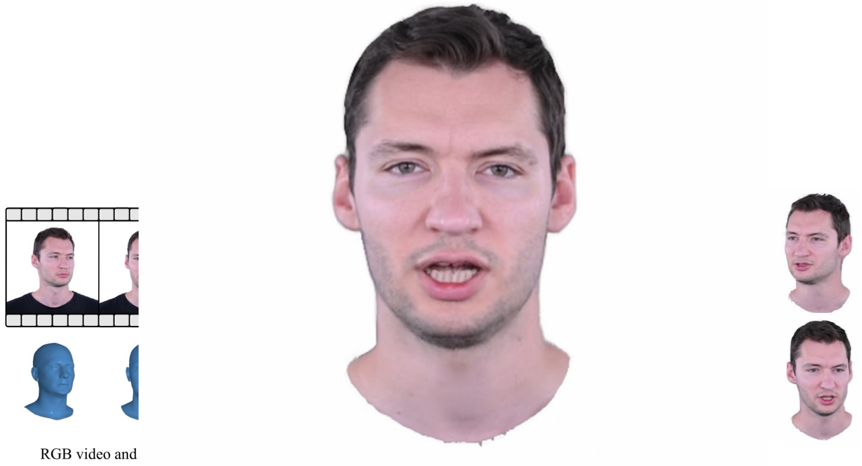
Take-home Messages of GEM

1. GEM distills heavy CNN-based avatars into efficient linear Eigenbases without needing 3DMM.
2. Quality and memory are adjustable via the number of bases.
3. Enables real-time image-based cross-reenactment with a pretrained ResNet.



All Projects During my PhD

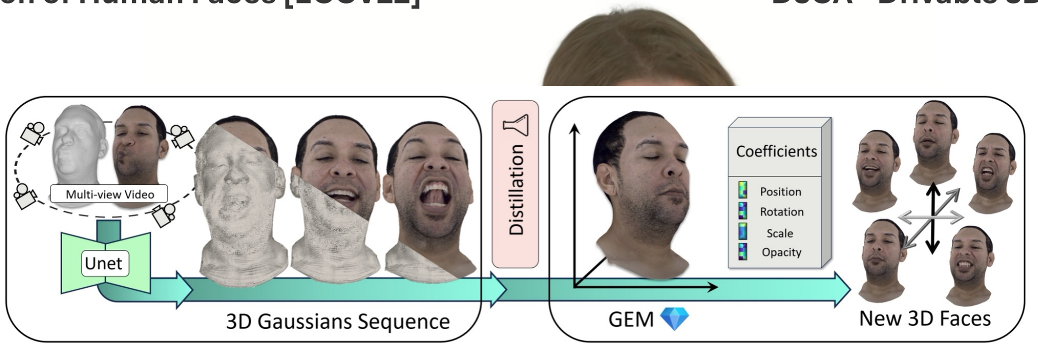
SynShot - Synthetic Prior for Few-Shot Drivable Head Avatar Inversion [CVPR25]



INSTA - Instant Volumetric Head Avatars [CVPR23]

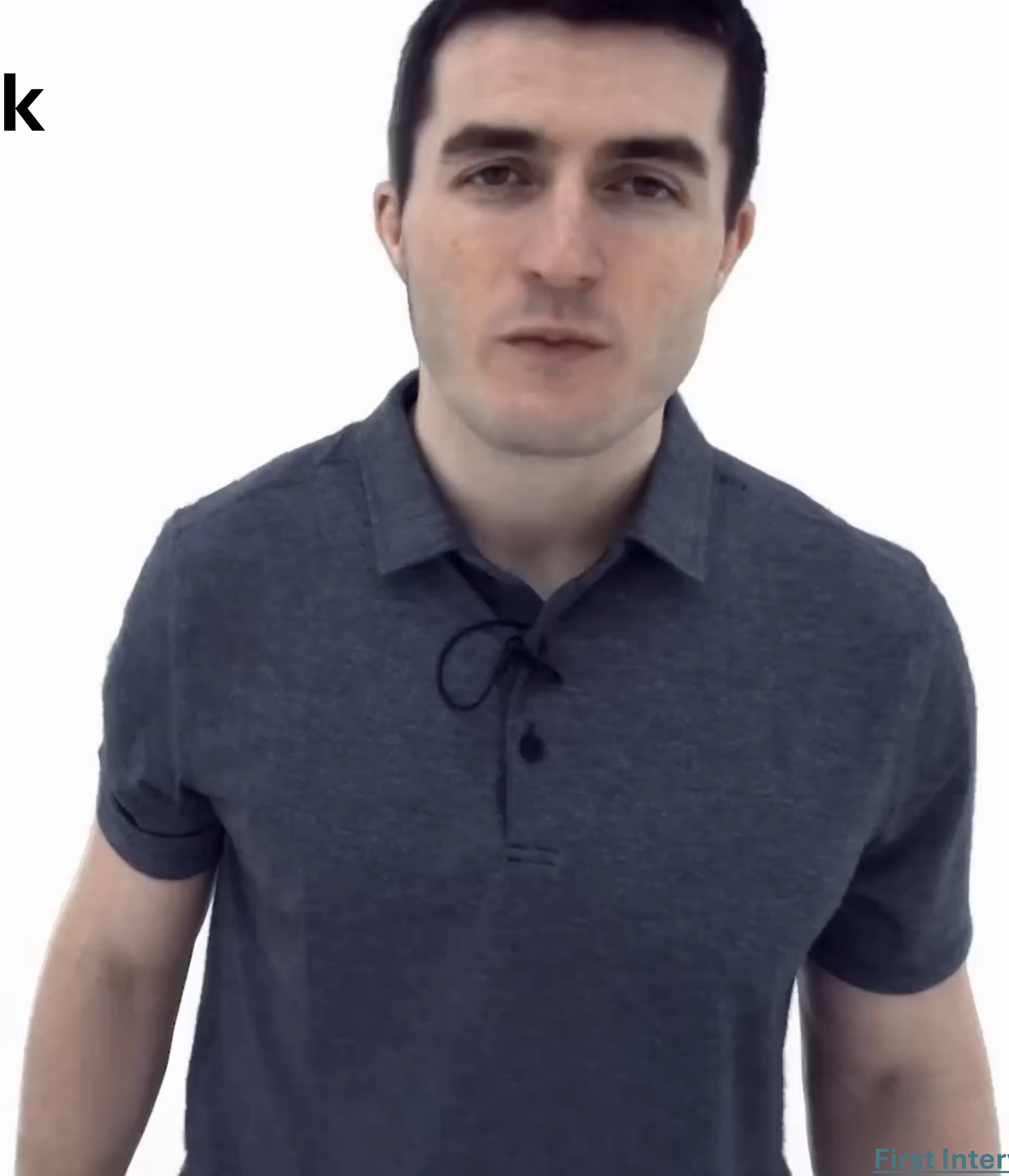
MICA - Towards Metrical Reconstruction of Human Faces [ECCV22]

D3GA - Drivable 3D Gaussian Avatars [3DV25]



GEM - Gaussian Eigen Models for Human Heads [CVPR25]

Future Work



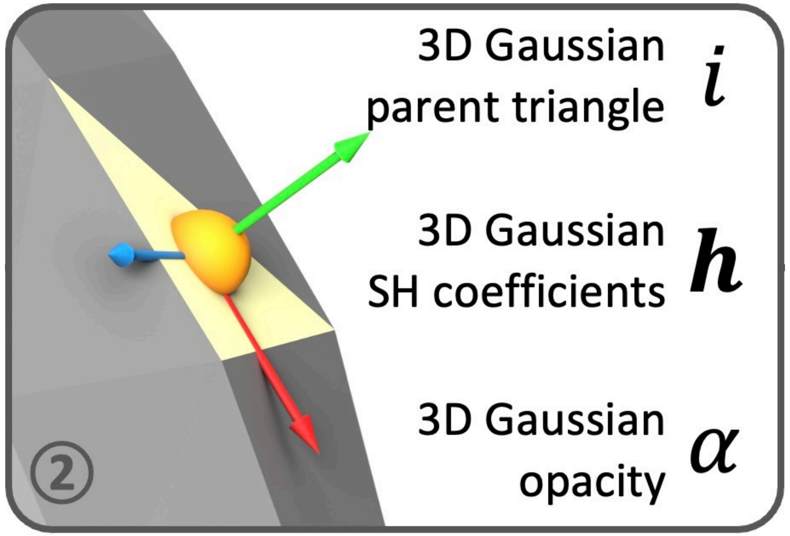
First Interview in the Metaverse | Mark and Lex | 2023


Thank you!

First Interview in the Metaverse | Mark and Lex | 2023

Additional Slides

(3DGS + FLAME)*



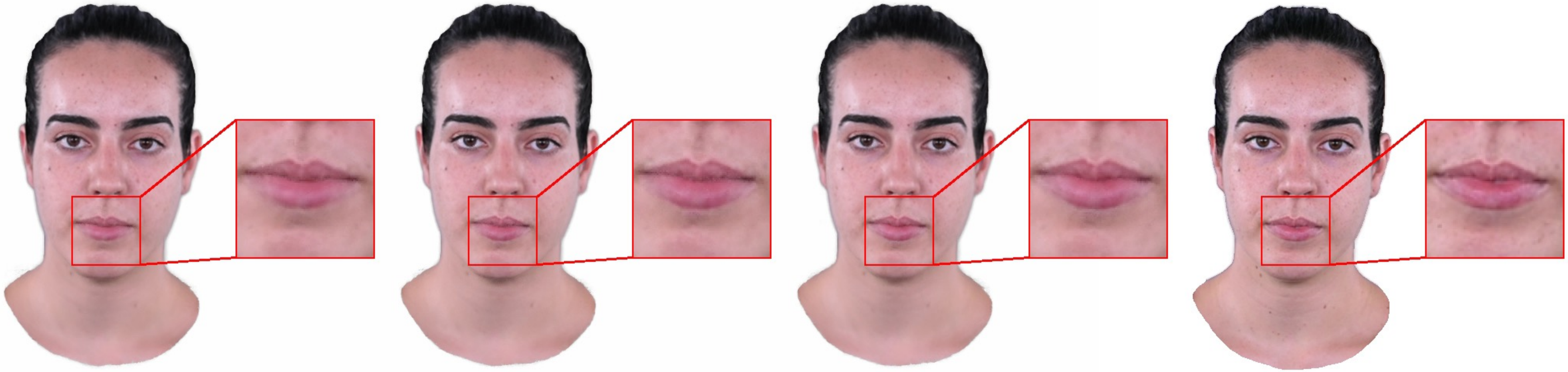
assign a 3D Gaussian at the center of each triangle

*Photorealistic Head Avatars with Rigged 3D Gaussians – Qian *et al.*

Results: Qualitative Comparison (Novel View)

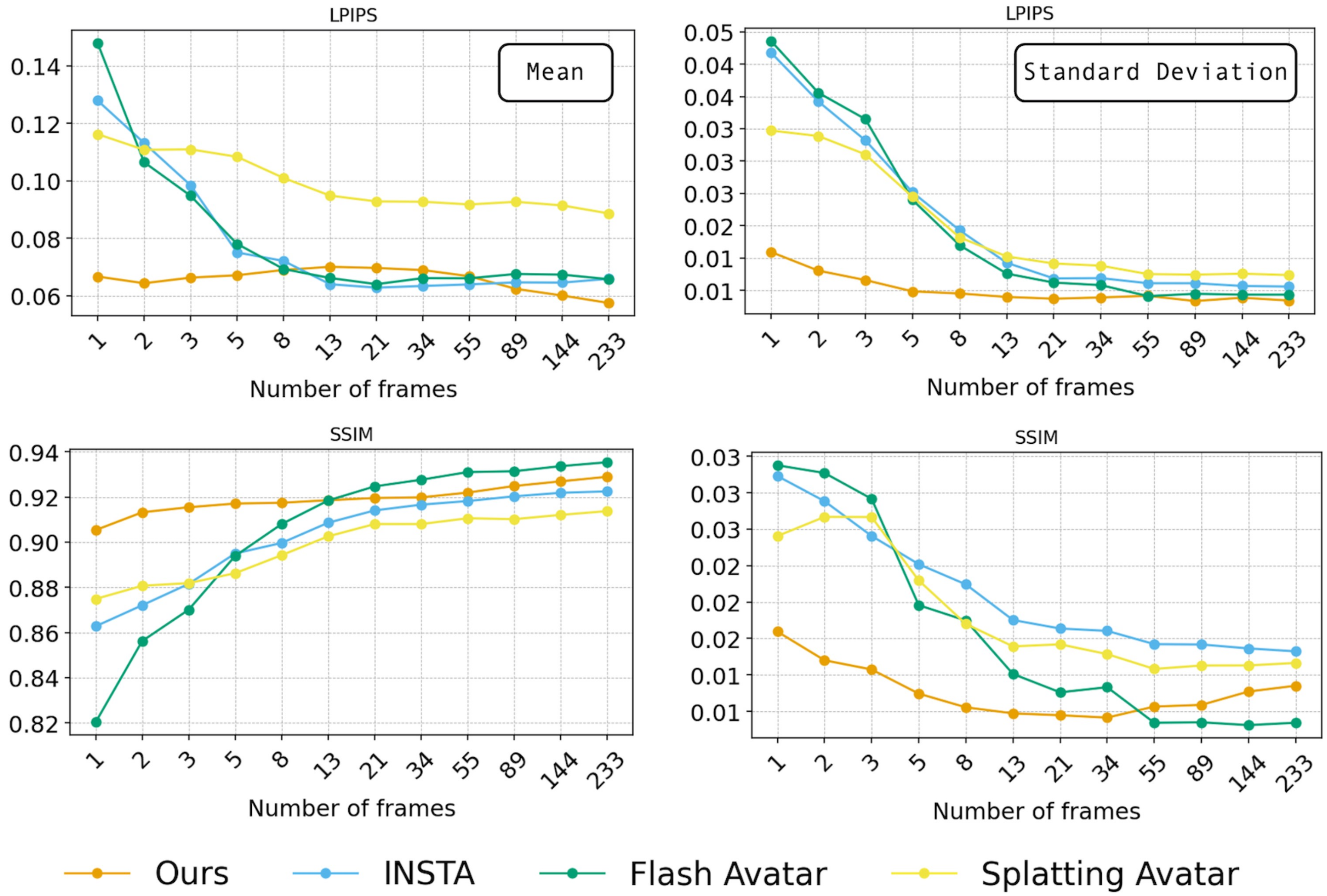
NHA [Grassal et al.] NeRFace [Gafni et al.] IMAvatar [Zheng et al.] Ours

Results: Ablation Studies



w/o mouth weigthing | w/o conditioning | Final | Ground Truth

Results: Quantitative Comparison



Results: Ablation Study (Architecture)

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

Results: Novel Expressions

Method	PSNR ↑	LPIPS ↓	SSIM ↑	L1 ↓
AG	29.0114	0.0812	0.9429	0.0099
GA	28.3137	0.0815	0.9433	0.0102
INSTA	27.9181	0.1153	0.9340	0.0128
Ours Net	29.2454	0.0777	0.9448	0.0096
Ours GEM	**32.6781**	**0.0675**	**0.9633**	**0.0069**

AG – Animatable Gaussians [Li et al.]
GA – Gaussian Avatars [Qian et al.]
INSTA – [Zielonka et al.]

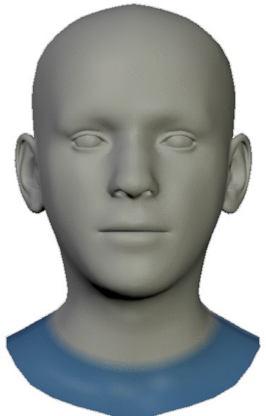
Results: Novel Viewpoint

Method	PSNR ↑	LPIPS ↓	SSIM ↑	L1 ↓
AG	32.4166	0.0712	0.9614	0.0066
GA	31.3197	0.0786	0.9567	0.0075
INSTA	27.7786	0.1232	0.9294	0.0163
Ours Net	32.4622	0.0713	0.9617	0.0067
Ours GEM	**33.5528**	**0.0678**	**0.9662**	**0.0061**

AG – Animatable Gaussians [Li et al.]
GA – Gaussian Avatars [Qian et al.]
INSTA – [Zielonka et al.]

GEM: Localized PCA

GEOMETRY/eyeballs COMP 00 = −3.0



GEM: Localized PCA

Source GEM Localized GEM