ACM SIGGRAPH / Eurographics Symposium on Computer Animation 2026
D. Levin and M. Chu
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 8

Multi-View Face and Gesture Animation with Dynamic Gaussians

Alireza Javanmardi[1†], Vippin Kumar Jeetmal[2†], Christen Millerdurai[1], Alain Pagani[1] and Didier Stricker[1,2]

[1]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)
[2] Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau (RPTU)
[†]Equal contribution

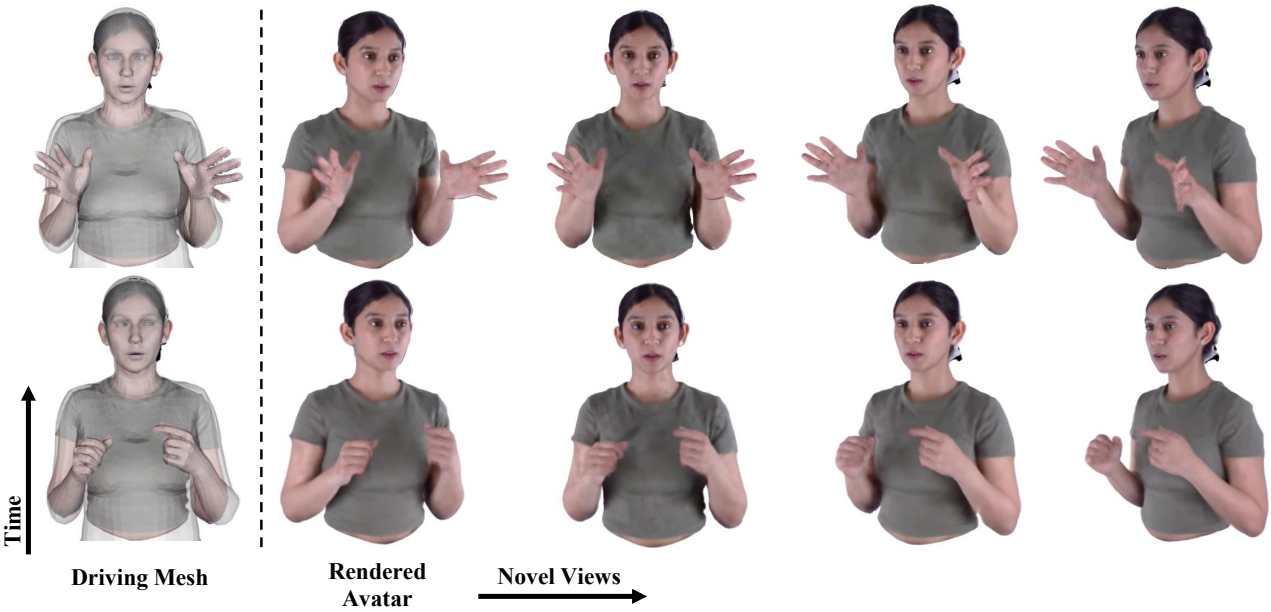

Figure 1: ***Multi-view face and gesture animation.*** *Given a time-varying driving mesh (left), our method (**MVFGA**) animates a reconstructed 3D avatar learned from multi-view data and synthesizes geometry-consistent novel-view renderings (right), faithfully reproducing facial expressions and hand gestures.*

Abstract

Creating photorealistic 3D human avatars with realistic upper-body motion remains challenging. Existing approaches either focus on the head and overlook hand gestures, or reconstruct the full body but fail to preserve fine-grained facial fidelity and hand pose accuracy. As a result, current methods struggle to capture the subtle dynamics of facial expressions and hand gestures that are crucial for natural human communication. While methods based on full-body parametric models enable avatar reconstruction from monocular or multi-view inputs, they often lack accurate facial animation and detailed hand articulation. To address these limitations, we propose MVFGA, a novel multi-view-consistent pipeline for generating realistic upper-body avatars. Our approach models the face and hands separately and fuses them with a parametric upper-body mesh model, enabling the capture of fine-grained facial expressions and hand poses for accurate upper-body avatar reconstruction. We then splat 3D Gaussians onto the obtained mesh, enabling high-quality rendering of dynamic avatars from novel viewpoints. Furthermore, we introduce MVFGA-MoCap, a multi-view upper-body motion capture dataset featuring controlled facial expression sequences, diverse hand gestures, and free-form communication. Experiments show that MVFGA generates visually realistic avatars with high-fidelity facial expressions and hand motions, outperforming baselines for upper-body avatar animation. Project page: https://dfki-av.github.io/MVFGA/

CCS Concepts

• ***Computing methodologies*** → *Animation;*

1. Introduction

Animating humans in photorealistic 3D has become an increasingly important research problem due to the growing demand for authentic and engaging digital interactions across applications such as virtual reality, gaming, remote communication, and content creation. Beyond entertainment, realistic human avatars are becoming essential in high-stakes scenarios such as remote medical training, surgical teleoperation, virtual education, and collaborative design, where subtle facial expressions and precise hand gestures play a critical role in effective communication [GJGP24; GJK*25]. In these settings, digital avatars act as proxies for real individuals in virtual environments and are often personalized to closely resemble the target user. However, imperfect appearance or motion can reduce user comfort rather than enhance immersion, a response commonly attributed to the uncanny valley effect [70], which describes negative reactions to near-human entities with subtle perceptual or behavioral flaws.

The main challenge is to disentangle the actor's appearance from their motion, including pose and facial expressions. This disentanglement enables animation at inference time using novel motions, either from new sequences of the same actor or transferred from a different actor. Consequently, recent research has largely followed two directions. One line of work focuses on animating facial expressions only, producing talking-head or portrait-level animations [WML21; MLW*24; KGN24]. Another line aims to generate holistic full-body animations [SWR*21; WYBD22; QGL*25]. While these approaches have achieved impressive results, they often fall short in capturing realistic upper-body communication, which critically involves both expressive facial dynamics and precise hand gestures. This limitation significantly reduces their applicability in scenarios where hand motion and facial expression are tightly coupled.

Recent progress in human animation has been advanced by both graphics-based [QKS*24; JSZ*25; ZLL*25] and generative approaches [TZTL25; BLW*24; TXH*25]. While these methods have shown promising performance in one-/few-shot settings, they still struggle to produce lifelike avatars with multi-view consistency. In practice, reconstructed avatars often suffer from degraded facial fidelity, incomplete finger articulation, and view-dependent artifacts, which become particularly noticeable when generative backbones inpaint unseen regions. These limitations highlight the need for a framework that can reconstruct and render reliable, high-quality avatars suitable for real-world applications.

To address these limitations, we introduce **Multi-View Face and Gesture Animation with Dynamic Gaussians (MVFGA)**, a multi-view-consistent framework for upper-body avatar reconstruction and reenactment as shown in Figure 1. Our end-to-end pipeline builds an animatable upper-body parametric model by integrating facial expressions and hand articulations into a unified representation, enabling detailed facial and hand motion. We then splat 3D Gaussians [KKLD23] onto the mesh to synthesize photorealistic dynamic avatars under novel viewpoints. The generated avatars can be driven from a monocular video while preserving multi-view-consistent appearance, enabling real-time and immersive applications. To support training and evaluation, we introduce **MVFGA-MoCap**, a comprehensive multi-view dataset for upper-body face and gesture animation. It includes synchronized videos of 15 subjects captured by 17 calibrated cameras from diverse viewpoints, covering controlled facial expressions, diverse hand gestures, and free-form communication. Our main contributions are summarized as follows:

- We propose an animatable upper-body parametric model that separately integrates face and hand parameterizations into a unified representation.
- We present a novel multi-view pipeline for reconstructing the upper-body motion and appearance of a human actor, enabling photorealistic upper-body avatar synthesis and novel-view rendering.
- We introduce **MVFGA-MoCap**, a multi-view upper-body motion-capture dataset featuring controlled facial expressions and diverse hand gestures.

2. Related Work

2.1. Graphics-based Avatar Animation

Early approaches typically rely on parametric face or body models, such as FLAME [LBB*17], or SMPL [LMR*15], where personalized geometry is reconstructed and animated using pose and expression parameters [TZS*16]. Once the geometry is reconstructed, appearance is modeled via texture mapping and rendered using rasterization-based pipelines. These methods offer strong structural consistency and real-time performance; however, their visual fidelity is often limited by the expressiveness of the rendering pipeline, making it challenging to capture fine-scale details such as hair, accessories, and subtle appearance variations.

Neural rendering techniques have significantly advanced realism by replacing explicit mesh-based representations with implicit neural representations. Neural Radiance Fields (NeRFs) [MST*20] and their dynamic variants enable high-quality, view-consistent avatar reconstruction and animation [GTZN21; PSH*21; ASS23]. While these approaches substantially improve visual quality compared to traditional graphics-based methods, their reliance on computationally expensive ray sampling results in slow rendering speeds, limiting their practicality for real-time applications and deployment-critical scenarios.

More recently, 3D Gaussian Splatting has emerged as an efficient alternative that combines high visual fidelity with fast rendering performance [KKLD23]. Several works leverage Gaussian representations for head avatar reconstruction and animation, achieving improved efficiency over NeRF-based methods [QKS*24; XCL*24; TRM*25; AWB*25]. Other approaches extend Gaussian splatting to full-body animation, producing coherent holistic renderings but often lacking fine-grained facial and hand details [PZK*24; QGL*25; MSS24]. Conversely, face-centric methods achieve high-quality facial rendering but neglect articulated hand gestures. Additionally, recent few-shot [ZGL*25] or single-view Gaussian-based approaches [HGY*25; ZLL*25] require large-scale training datasets and often exhibit limited generalization to novel viewpoints. In contrast, we bridge the gap between face-centric and full-body methods by enabling

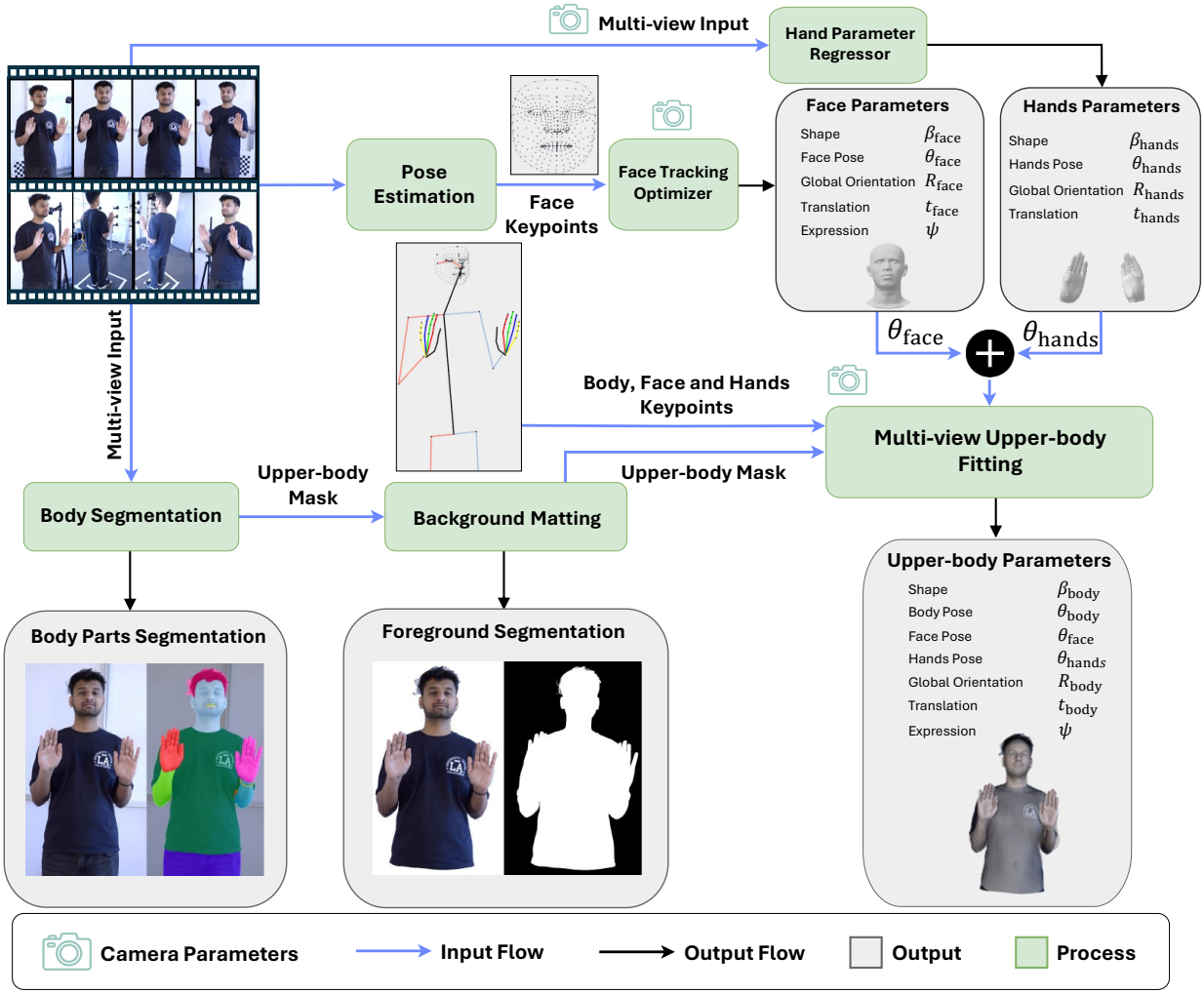

Figure 2: ***Upper-body mesh generation pipeline****: We perform keypoint extraction, body-part segmentation, and background matting, followed by refined face and hand fitting. The resulting face, hand, and body parameters are fused to obtain a complete upper-body representation with accurate shape, pose, and global translation.*

multi-view–consistent upper-body reenactment with expressive facial dynamics and articulated hand gestures.

2.2. Generative Avatar Animation

Generative-based avatar animation methods aim to synthesize animated faces or bodies directly from data, without explicit 3D reconstruction. Early approaches predominantly adopt adversarial learning frameworks, where motion is represented using 2D or 3D keypoints extracted from driving frames and used to warp source features for animation [SLT*19; JPS24; WML21; SWR*21]. While effective for talking-head synthesis, these methods often struggle with large pose variations and full-body motion due to the absence of explicit 3D structure.

More recent works explore latent-space manipulation via GAN inversion [WYBD22; WYBD24] or leverage diffusion models to improve visual fidelity and controllability [XZL*24; Hu24; ZCD*24; TXH*25; ZGW*24]. Diffusion-based approaches benefit from strong image priors and support diverse conditioning signals such as pose, motion, and audio. However, they typically require substantial computational resources and slow sampling procedures, limiting real-time applicability. Moreover, achieving long-term temporal consistency remains challenging, and the resulting avatars are not explicitly animatable in 3D. A recent line of work combines generative models with 3D representations such as Gaussian Splatting to enable few-shot 3D avatar animation [TZTL25; KRS*25]. Despite promising results, these approaches often suffer from hallucination of unseen regions, hindering consistent 3D avatar creation and reliable retargeting. In contrast, we adopt a graphics-based, explicitly animatable representation for improved controllability and efficiency.

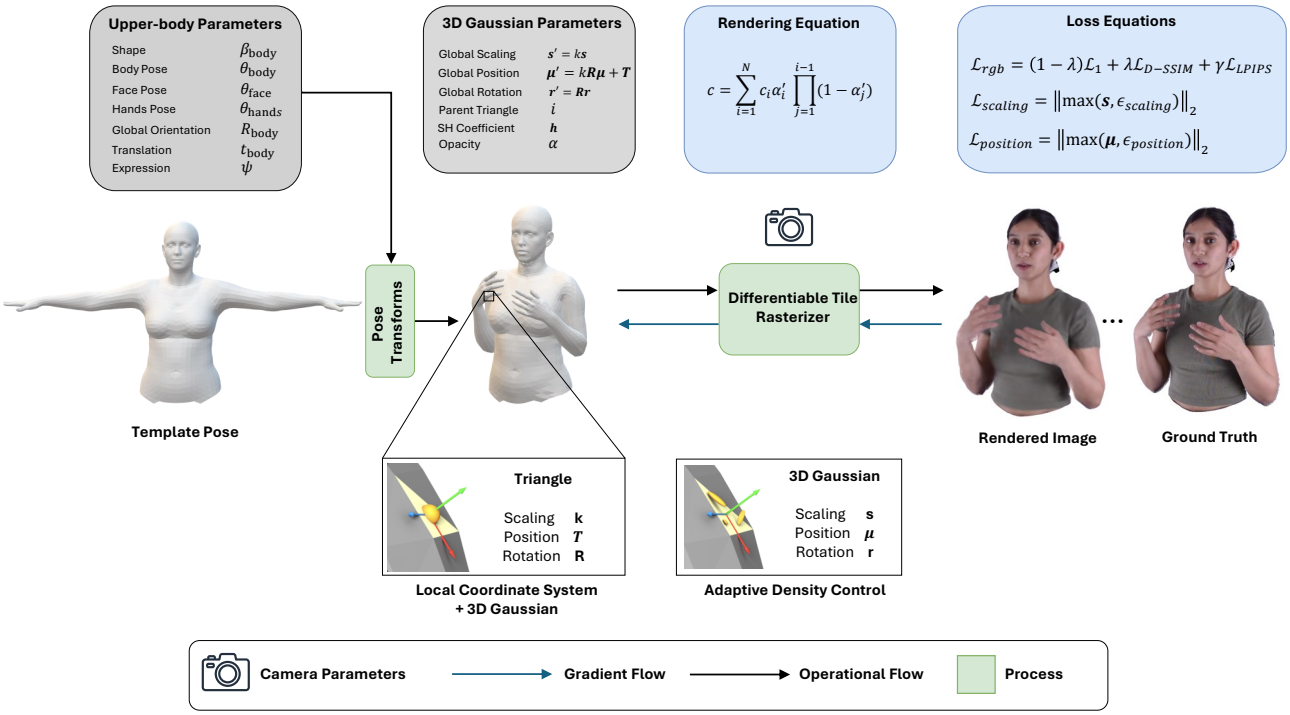

Figure 3: ***Overview of our avatar synthesis pipeline****. Given multi-view images and corresponding upper-body parameters, following GaussianAvatar [QKS*24] the template mesh is posed into the deformed space. Each triangle is then assigned a 3D Gaussian representation. These Gaussians are rasterized using a tile-based rasterizer to produce rendered images, which are supervised using an RGB reconstruction loss. An adaptive density control mechanism is employed during training to dynamically densify or prune the Gaussians for efficient and accurate representation.*

3. Methodology

Our aim is to create and animate a realistic 3D upper-body avatar of an actor that maintains high fidelity under novel viewpoints. First, we capture the actor using a synchronized multi-view rig and fit our upper-body parametric model to the recordings, yielding a temporally consistent mesh that deforms with the actor's motion (Sec. 3.1). Next, we place 3D Gaussians on the mesh surface and optimize them to reconstruct the actor's appearance across time and camera viewpoints (Sec. 3.3). This process produces a deformable Gaussian-splat representation, which can be animated at inference time under novel target motions and viewpoints. Finally, we describe the dataset collected during this process in Sec. 3.4.

3.1. Upper-body Mesh Reconstruction

As shown in Figure 2, we reconstruct a detailed 3D upper-body mesh from temporally aligned multi-view RGB video captured by 17 calibrated cameras. Our pipeline begins by extracting 2D full-body keypoints in all views to obtain a robust pose initialization and reduce ambiguity caused by self-occlusions. To capture expressive motion, we further refine regions where whole-body parametric fitting is less accurate. For the face, we estimate facial shape and expression parameters from the video using an optimization-based face tracking procedure. For the hands, we process the left and right hand regions independently using a transformer-based estimator that regresses hand pose and shape parameters. Specifically, we use MICA [ZBT22] for face parameters and HaMeR [PSR*24] to estimate MANO [RTB17] hand parameters. A key challenge is that these methods are designed and optimized for monocular inputs, and therefore do not directly account for our calibrated multi-view camera intrinsics and extrinsics, nor do they produce view-consistent parameter estimates. To address this, we generate multiple candidate estimates and select the one with the lowest aggregated multi-view 2D reprojection error, while rejecting outliers. Next, we define our upper-body parametric model, a novel extension of SMPL-X [PCG*19] that restricts the full-body mesh topology to the upper body by removing vertices and faces that lie outside the upper-body region. We preserve its parameterization, including the shape space as well as pose and expression parameters, making our model backward compatible with SMPL-X. We fit this upper-body model to the actor by jointly optimizing the body parameters using the multi-view 2D keypoints together with the estimated face and hand parameters, while explicitly accounting for the calibrated camera intrinsics and extrinsics (Multi-view Upper-body Fitting in Figure 2). This yields a multi-view-consistent upper-body mesh with highly articulated face and hand regions (see Figure 4). Finally, we apply background matting and semantic body-part

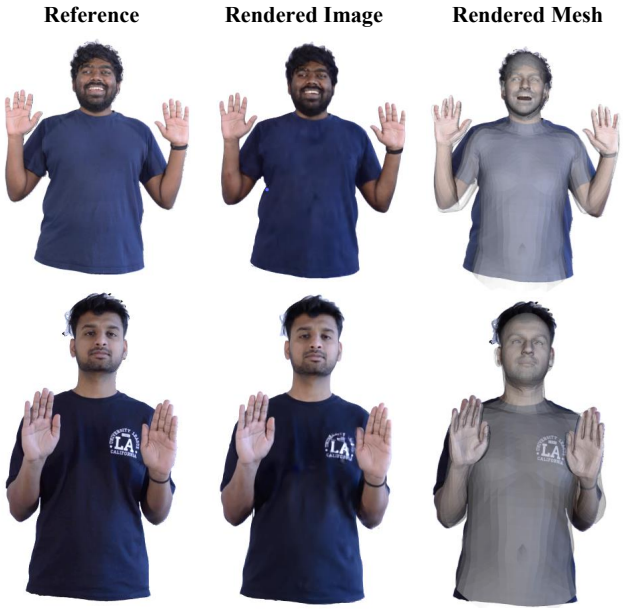

Figure 4: ***Upper-body mesh fitting****. From left to right: reference, rendered image, and rendered fitted mesh.*

segmentation to isolate the actor and generate a clean upper-body mask. This mask is used to optimize the 3D Gaussians and provides a high-fidelity matte for the rendered avatar. Additional details are provided in the supplementary material.

3.2. 3D Gaussian Splatting Preliminary

We build on 3D Gaussian Splatting (3DGS) [KKLD23], which reconstructs a scene from multi-view images and calibrated cameras using anisotropic 3D Gaussians. Each splat is centered at a mean μ and represented by a covariance matrix Σ, defining the density

$$G(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right). \tag{1}$$

To ensure Σ remains positive semi-definite during optimization, 3DGS parameterizes each Gaussian as an oriented ellipsoid with rotation $\mathbf{R}$ and scaling $\mathbf{S}$:

$$\Sigma = \mathbf{R}\mathbf{S}\mathbf{S}^\top \mathbf{R}^\top. \tag{2}$$

In practice, each ellipsoid is stored as $\mu \in \mathbb{R}^3$, a scaling vector $\mathbf{s} \in \mathbb{R}^3$, and a quaternion $\mathbf{q} \in \mathbb{R}^4$ (we denote its rotation matrix by $\mathbf{r} \in \mathbb{R}^{3\times3}$). For rendering, Gaussians are projected to the image plane and composited in depth order using alpha blending:

$$\mathbf{C} = \sum_{i=1}^{N} \mathbf{c}_i \alpha_i' \prod_{j=1}^{i-1} (1-\alpha_j'), \tag{3}$$

where $\mathbf{c}_i$ is modeled with low-order spherical harmonics and α_i' is obtained from the projected 2D Gaussian weighted by opacity.

3.3. Upper-body Avatar Animation Pipeline

Our animation pipeline, as illustrated in Figure 3, shows the upper-body mesh augmented with a set of surface-attached 3D Gaussians. Given the posed mesh and calibrated cameras, we render the Gaussians via a differentiable rasterizer and optimize their parameters to match the observed multi-view appearance. To maintain both efficiency and fidelity, we employ adaptive density control throughout training.

3.3.1. Initialization.

Following GaussianAvatar [QKS*24], we adapt their 3D Gaussian initialization strategy to place 3D Gaussians on our upper-body parametric mesh model. Specifically, each triangle of the mesh is initialized with a single 3D Gaussian splat positioned at the triangle center. Each Gaussian is parameterized in the local coordinate system of its parent triangle by a mean location μ, a rotation matrix $\mathbf{r}$, and a scaling vector $\mathbf{s}$. At initialization, μ is set to the local origin, $\mathbf{r}$ is the identity rotation, and $\mathbf{s}$ is set to a unit scale. During rendering, these local parameters are transformed into the global coordinate system using:

$$\mathbf{r}' = \mathbf{R}\mathbf{r}, \tag{4}$$

$$\mu' = k\mathbf{R}\mu + \mathbf{T}, \tag{5}$$

$$\mathbf{s}' = k\mathbf{s}, \tag{6}$$

where $\mathbf{R}$ and $\mathbf{T}$ denote the global rotation and translation of the corresponding mesh triangle, respectively, and k is a scalar scale factor that describes the local-to-global scaling of the triangle.

3.3.2. Adaptive Density Control

To capture high-frequency appearance details that are not explicitly modeled by the mesh geometry, we employ an adaptive density control mechanism based on [KKLD23; QKS*24] that dynamically adds or removes splats based on view-space positional gradients and opacity statistics. When a Gaussian is split or cloned during densification, the newly created splats inherit the same parent triangle as the original, ensuring that they remain consistently attached to the mesh surface. This association is maintained by storing the index of the parent triangle for each Gaussian. In addition, pruning operations remove splats with persistently low opacity to improve computational efficiency. To avoid artifacts in frequently occluded regions (e.g., the eyes), we enforce that each mesh triangle always retains at least one associated Gaussian splat, even after pruning.

3.3.3. Optimization Objectives.

The 3D Gaussian splats are rendered into RGB images using a differentiable tile-based rasterizer and supervised using a combination of pixel-wise and perceptual losses. The RGB reconstruction loss is defined as:

$$\mathcal{L}_{\text{rgb}} = (1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\text{D-SSIM}} + \gamma\mathcal{L}_{\text{LPIPS}}, \tag{7}$$

where λ and γ are loss weighting terms, $\mathcal{L}_1$ denotes the ℓ_1 loss, $\mathcal{L}_{\text{D-SSIM}}$ is the differentiable SSIM loss, and $\mathcal{L}_{\text{LPIPS}}$ is the perceptual loss as suggested in [HFW*24].

In addition, we incorporate regularization terms to ensure stable

training. A position regularization loss constrains Gaussian means to remain close to their parent triangle:

$$\mathcal{L}_{\text{position}} = \left\| \max(\mu, \epsilon_{\text{position}}) \right\|_2, \tag{8}$$

where $\epsilon_{\text{position}} = 1$ allows small deviations due to triangle scaling. We further introduce a scaling regularization loss to prevent splats from becoming excessively large relative to their parent triangle:

$$\mathcal{L}_{\text{scaling}} = \left\| \max(\mathbf{s}, \epsilon_{\text{scaling}}) \right\|_2, \tag{9}$$

where $\epsilon_{\text{scaling}} = 0.6$ disables this penalty for sufficiently small splats. The final objective is:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{position}} \mathcal{L}_{\text{position}} + \lambda_{\text{scaling}} \mathcal{L}_{\text{scaling}}. \tag{10}$$

Here, $\lambda_{\text{position}}$ and λ_{scaling} are weighting coefficients for the position and scaling regularizers, respectively. We apply these regularization terms only when the RGB reconstruction loss is active for the corresponding timestep.

3.4. Dataset

We collect a high-quality multi-view upper-body motion capture dataset targeting detailed facial expressions, hand articulation, and torso motion. Data are recorded in a calibrated 17-camera RGB studio setup, including 15 front-facing cameras spanning a 150° arc and two rear cameras for improved body keypoint detection. All videos are synchronized and captured at 1920×1080 resolution and 25 FPS under uniform studio lighting. The dataset contains 15 participants (8 male, 7 female; ages 24–32) performing diverse facial expressions and two-handed gestures. Further details on capture, calibration, statistics, and post-processing are provided in the supplementary material.

4. Experiments and Results

4.1. Experimental Setup

We evaluate our framework on our multi-view upper-body dataset under three settings: (1) *self-reenactment*, where we drive an avatar using a held-out sequence of the same subject with unseen poses and expressions and render the frontal view; (2) *novel-view synthesis*, where we animate the avatar using motions from training sequences and render from a held-out camera viewpoint; and (3) *cross-identity animation*, where we transfer poses and expressions from one subject to animate the avatar of another subject.

4.2. Implementation Details

Our pipeline begins by processing multi-view video to obtain upper-body representation. We extract the foreground subject using BiRefNet [ZGF*24] and segment semantic body parts with Sapiens [KBM*24]. Initial 2D keypoints from MediaPipe [LTN*19] are combined with detailed FLAME [LBB*17] face parameters (from MICA [ZBT22]) and MANO [RTB17] hand parameters (from HaMeR [PSR*24]) within the EasyMocap [DFJ*21] framework to produce temporally consistent SMPL-X [PCG*19] parameters. We then derive our upper-body parametric model by selecting the upper-body vertices using the SMPL-X part-segmentation map and discarding vertices and triangles outside the upper-body region. During optimization of the Gaussians, we use the semantic part segmentation to mask out lower-body pixels in each view. All parameters are optimized with Adam [KB17], using the original learning rates for the Gaussians. The SMPL-X parameters are fine-tuned with component-specific learning rates: 1×10^{-8} (global translation), 1×10^{-4} (facial expressions), and 1×10^{-6} (body pose). We set loss weights $\lambda = 0.2$, $\gamma = 0.04$, $\lambda_{\text{position}} = 0.01$, and $\lambda_{\text{scaling}} = 1$. Each avatar is optimized for 600K iterations on a single NVIDIA H100 GPU, taking approximately 8 hours.

4.3. Baselines

For comparison, we evaluate state-of-the-art methods for upper-body reenactment. We include recent generative approaches, AnimateAnyone [Hu24], MagicAnimate [XZL*24], and Champ [ZCD*24]. We also compare against the graphics-based baseline GUAVA [ZLL*25], which explicitly supports upper-body avatar modeling. Most methods primarily target either face-only animation [XCL*24; QKS*24] or full-body synthesis [QGL*25; LWP*24] and do not perform reliably on our upper-body setting.

4.4. Metrics

We evaluate image quality using L1 error, Peak Signal-to-Noise Ratio (PSNR) [HZ10], Structural Similarity Index Measure (SSIM) [WBSS04], and Learned Perceptual Image Patch Similarity (LPIPS). All metrics are computed using the Disco evaluation toolkit [WLL*24]. For pose accuracy, we report Average Keypoint Distance (AKD) [Gas11], computed on MediaPipe landmarks [LTN*19] for the face, hands, and torso. To assess facial identity preservation, we compute cosine similarity (CSIM) using ArcFace features [DGXZ19]. To evaluate temporal consistency, particularly for long-form animation rendering, we use Temporal Jittering Error (TJE) [JJH*26].

4.5. Evaluation

Protocol. For comparison methods, we use the first frame of each sequence as the source image and the remaining frames as the driving video, enabling frame-wise comparison across methods. As several generative baselines struggle with long video synthesis, we split each sequence into 100-frame clips for evaluation. We report results on four held-out identities.

4.5.1. Quantitative Results

We report quantitative results using the metrics described in Sec. 4.4 in Table 1. Our method achieves the best overall performance across all metrics, indicating strong detail preservation without introducing artifacts. In particular, we obtain an LPIPS of 0.053 and a PSNR of 25.93, while Champ ranks second in LPIPS and GUAVA ranks second in SSIM and PSNR. We also achieve an SSIM of 0.938 and an L1 error of 4.06, compared to 9.59 for AnimateAnyone (second best).

For pose accuracy, Table 2 reports Average Keypoint Distance (AKD) for face, hands, and torso. Our method achieves the lowest AKD on hands and torso, and remains competitive on face AKD

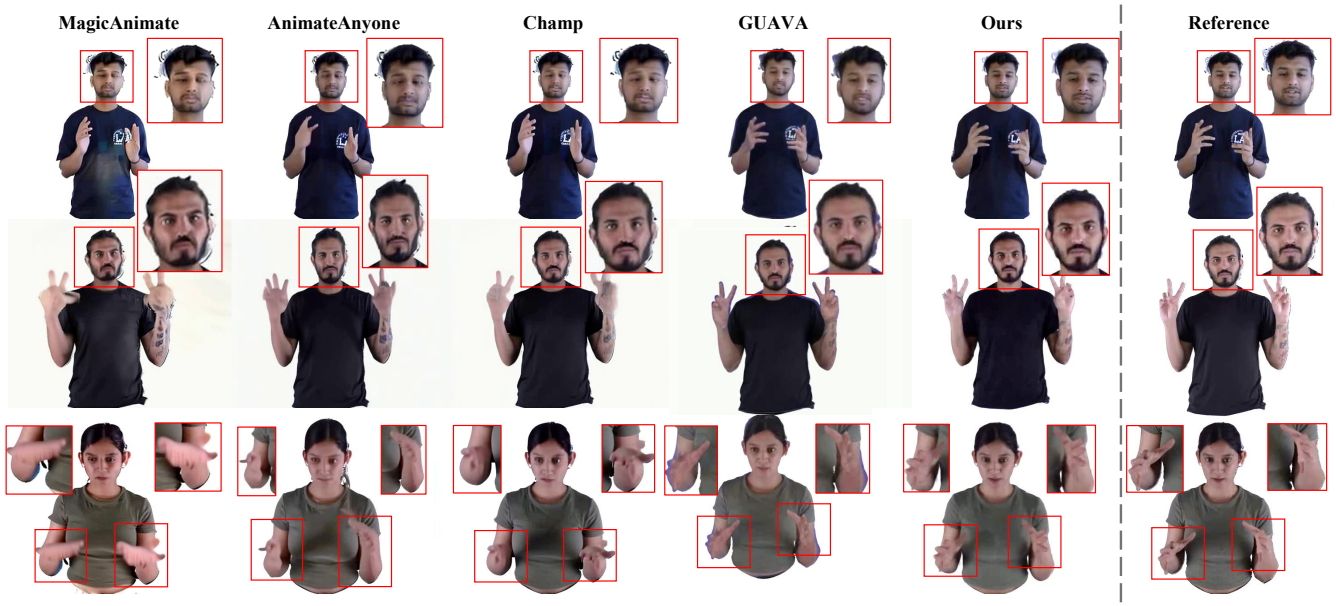

Figure 5: ***Qualitative comparison***. *Our method captures facial expressions, hand gestures, and upper-body poses across diverse motions while better preserving appearance compared to MagicAnimate[XZL*24], AnimateAnyone [Hu24], Champ [ZCD*24], and GUAVA [ZLL*25].*

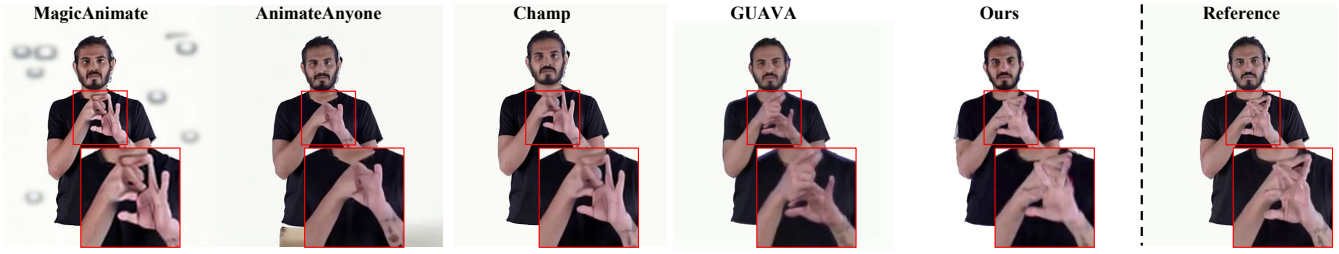

Figure 6: ***Qualitative comparison under complex hand articulation***. *The highlighted region contains a challenging gesture with closely interacting fingers and strong self-occlusions. Compared with prior methods, our approach better preserves finger structure, hand pose accuracy, and overall appearance consistency.*

(0.18), close to GUAVA (0.15). Finally, we obtain the highest CSIM score of 0.85, demonstrating strong facial identity preservation. We evaluate temporal consistency using the Temporal Jittering Error (TJE) [JJH*26] over 1k-frame intervals. As shown in Figure 9, diffusion-based methods exhibit higher jitter due to their stochastic nature, with performance degrading as sequence length increases. While recent approaches incorporate motion modules that can produce stable results for short clips, they still face limitations in GPU memory requirements and struggle to maintain temporal consistency over longer sequences. In contrast, graphics-based approaches, including GUAVA and ours, achieve lower TJE and maintain stable temporal dynamics across extended durations.

4.5.2. Qualitative Results

Self-reenactment. As shown in Figure 5, MagicAnimate, AnimateAnyone, and Champ often introduce background artifacts and fail to reproduce fine-grained hand gestures, while also exhibiting weaker identity preservation. In contrast, both GUAVA and our method generate sharper renderings with accurate pose and expression. However, GUAVA frequently loses facial identity details, whereas our approach preserves identity more faithfully and produces more photorealistic results. We further observe that our method remains robust under challenging hand articulations involving severe finger self-occlusions and fine-grained finger

Table 1: ***Quantitative comparison on our captured multi-view human upper-body dataset.*** *Best results are in* ***bold*** *and second best are underlined.*

Method	L1 ↓	SSIM ↑	PSNR ↑	LPIPS ↓
MagicAnimate [XZL*24]	17.61	0.823	14.90	0.171
AnimateAnyone [Hu24]	9.59	0.859	19.62	0.132
Champ [ZCD*24]	11.07	0.890	19.36	0.102
GUAVA [ZLL*25]	15.24	0.897	22.33	0.118
Ours	**4.06**	**0.938**	**25.93**	**0.053**

Figure 7: ***Novel-view synthesis.*** *Rendering at unseen yaw angles ($+5^\circ$, $+20^\circ$). Our method preserves geometry and appearance better than GUAVA [ZLL*25].*

Table 2: ***Quantitative comparison using keypoint and identity metrics.*** *AKD denotes Average Keypoint Distance for face (F), hands (H), and Torso (T).*

Method	AKD_F ↓	AKD_H ↓	AKD_T ↓	CSIM ↑
MagicAnimate [XZL*24]	0.82	2.81	4.92	0.35
AnimateAnyone [Hu24]	0.37	1.98	6.14	0.54
Champ [ZCD*24]	0.41	2.27	5.14	0.51
GUAVA [ZLL*25]	**0.15**	0.76	1.39	0.82
Ours	0.18	**0.71**	**1.37**	**0.85**

interactions. As illustrated in Figure 6, competing methods often produce distorted finger configurations or fail to preserve the intended gesture, whereas our approach accurately reconstructs the hand shape while maintaining appearance consistency.

Novel-view synthesis. Since novel-view rendering is essential for 3D applications, we additionally compare against GUAVA, as the other baselines do not support view extrapolation. In Figure 7, we fix pitch and roll and vary the yaw angle to $+5^\circ$ and $+20^\circ$. Our method maintains consistent geometry and high-quality appearance under viewpoint changes, while GUAVA degrades as the viewpoint moves further away from the training cameras.

Cross-identity animation. We further evaluate cross-identity animation by transferring motion from one subject to another. As shown in Figure 8, our method preserves the driving gestures while maintaining the target identity.

4.6. Ablation Studies

We conduct ablation studies to quantify the impact of key components in our avatar synthesis pipeline. Table 3 reports results on a held-out subject under both self-reenactment and novel-view synthesis settings, allowing us to evaluate both motion reenactment quality and cross-view generalization.

4.6.1. Without FLAME and MANO.

We remove FLAME and MANO and optimize only the upper-body parameters (body pose, global rotation, and translation). This leads to a clear drop across all metrics in both settings. Qualitatively, the reconstructed avatars exhibit less accurate facial expressions and degraded hand articulation, which noticeably reduces realism and identity consistency. These results confirm that explicit face and hand modeling is crucial for high-fidelity avatar reconstruction and reenactment.

4.6.2. Without upper-body parametric fine-tuning.

We disable upper-body parametric fine-tuning during training and rely solely on the pre-fitted parameters obtained from multi-view keypoints and FLAME/MANO fits. Performance decreases in

Table 3: ***Ablation study on Novel-View synthesis and Self-Reenactment.*** *"Upp. Body FT" denotes fine-tuning upper-body mesh parameters. Best results are shown in* ***bold****.*

	Novel-View				Self-Reenactment			
Method	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	**2.46**	**24.94**	**0.958**	**0.074**	**2.71**	24.18	0.953	**0.075**
w/o FLAME & MANO	3.08	24.15	0.947	0.083	3.08	23.35	0.943	0.082
w/o Upp. BODY FT	2.67	24.10	0.953	0.080	2.74	**24.24**	**0.954**	0.077
w/o LPIPS	4.90	24.25	0.943	0.085	4.54	23.17	0.953	0.098

Figure 8: ***Cross-identity reenactment.*** *Motions from a driving subject are transferred to a target avatar while preserving identity and gestures.*

both settings, indicating that jointly refining the parametric model together with Gaussian appearance is important for accurate reconstruction.

4.6.3. Without LPIPS loss.

Finally, we remove the LPIPS term from the RGB reconstruction objective. This leads to lower perceptual quality, showing that

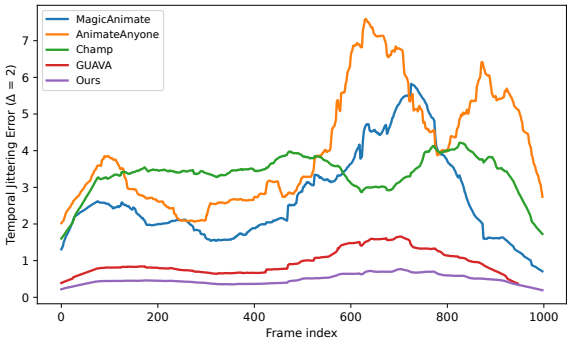

Figure 9: ***Temporal Jittering Error (TJE).*** *Lower values indicate better temporal consistency. TJE measures the discrepancy between motion differences in real and generated videos, capturing subtle jitter and flickering artifacts across frames.*

LPIPS provides complementary supervision beyond pixel-wise losses and improves overall visual fidelity.

4.7. Ethical Considerations and Potential Misuse.

Our method enables high-fidelity human avatar reconstruction and cross-identity animation, which can be misused for impersonation, deceptive media generation, or non-consensual content. We emphasize that such uses are unethical and strongly discourage them. The dataset used in this work contains identifiable facial and motion data; all participants provided informed consent for data capture and research use, following applicable institutional guidelines. We recommend that future releases of data or models include appropriate safeguards, such as usage restrictions, consent-based data sharing, and disclosure or watermarking mechanisms, to mitigate potential misuse and promote responsible deployment.

4.8. Limitations

Despite achieving consistent avatar reconstruction and reenactment, our approach still relies on a parametric model. While this representation is stable and easy to tune, it is difficult to obtain. In addition, full 360° rendering remains challenging with the current dataset and pipeline, as we restrict training and evaluation to frontal views where most appearance cues are visible. Moreover, our model does not explicitly model secondary motion

(e.g., clothing and accessories deformation), which can reduce realism under fast or complex movements.

5. Conclusion

We presented **MVFGA**, a multi-view framework for high-fidelity upper-body avatars with fine-grained facial expressions and articulated hand motion. Our key idea is to build a unified upper-body avatar representation by explicitly integrating detailed face and hand parameterizations into an upper-body parametric model, and coupling it with a deformable 3D Gaussian appearance field for photorealistic novel-view rendering. We evaluate MVFGA on our captured multi-view upper-body dataset under self-reenactment, novel-view synthesis, and cross-identity animation. MVFGA consistently outperforms strong generative and graphics-based baselines across image-quality metrics, producing sharper renderings with better identity preservation and more accurate facial and hand motion. Keypoint-based evaluation shows competitive pose accuracy, and ablations verify the importance of explicit FLAME and MANO integration, upper-body parametric fine-tuning during Gaussian optimization, and LPIPS supervision. We also introduce **MVFGA-MoCap**, a synchronized, calibrated multi-view dataset with fitted parametric models to support future research on upper-body avatar reconstruction and animation. We hope this research will support continued progress on high-fidelity upper-body avatars and encourage further exploration of multi-view datasets and representation.

6. Acknowledgments

This work was partially funded by the Horizon Europe programme under the project IRIS-XR, Grant Agreement No. 101298672.

References

[70] "Bukimi no tani [The uncanny valley]." *Energy* 7 (1970), 33 2.

[ASS23] ATHAR, SHAHRUKH, SHU, ZHIXIN, and SAMARAS, DIMITRIS. "Flame-in-nerf: Neural control of radiance fields for free view face animation". *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 2023, 1–8 2.

[AWB*25] ANEJA, SHIVANGI, WEISS, SEBASTIAN, BAEZA, IRENE, et al. "Scaffoldavatar: High-fidelity gaussian avatars with patch expressions". *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 2025, 1–11 2.

[BLW*24] BUEHLER, MARCEL C., LI, GENGYAN, WOOD, ERROLL, et al. "Cafca: High-quality Novel View Synthesis of Expressive Faces from Casual Few-shot Captures". *ACM SIGGRAPH Asia 2024 Conference Paper*. 2024. DOI: 10.1145/3680528.3687580 2.

[DFJ*21] DONG, JUNTING, FANG, QI, JIANG, WEN, et al. "Fast and Robust Multi-Person 3D Pose Estimation and Tracking from Multiple Views". *T-PAMI*. 2021 6.

[DGXZ19] DENG, JIANKANG, GUO, JIA, XUE, NIANNAN, and ZAFEIRIOU, STEFANOS. "Arcface: Additive angular margin loss for deep face recognition". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, 4690–4699 6.

[Gas11] GASHLER, MICHAEL. "Waffles: A Machine Learning Toolkit". *Journal of Machine Learning Research* 12.69 (2011), 2383–2387 6.

[GJGP24] GASCH, CRISTINA, JAVANMARDI, ALIREZA, GARCIA-PALACIOS, AZUCENA, and PAGANI, ALAIN. "Avatar quality: A study on presence and user preference". *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*. 2024, 102–108. DOI: 10.1109/AIxVR59861.2024.00022 2.

[GJK*25] GASCH, CRISTINA, JAVANMARDI, ALIREZA, KHAN, AMEER, et al. "Exploring Avatar Utilization in Workplace and Educational Environments: A Study on User Acceptance, Preferences, and Technostress". *Applied Sciences* 15.6 (2025). ISSN: 2076-3417. DOI: 10.3390/app15063290. URL: https://www.mdpi.com/2076-3417/15/6/3290 2.

[GTZN21] GAFNI, GUY, THIES, JUSTUS, ZOLLHÖFER, MICHAEL, and NIESSNER, MATTHIAS. "Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, 8649–8658 2.

[HFW*24] HU, HEZHEN, FAN, ZHIWEN, WU, TIANHAO, et al. "Expressive gaussian human avatars from monocular rgb video". *Advances in Neural Information Processing Systems* 37 (2024), 5646–5660 5.

[HGY*25] HE, YISHENG, GU, XIAODONG, YE, XIAODAN, et al. "LAM: Large Avatar Model for One-shot Animatable Gaussian Head". *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 2025, 1–13 2.

[Hu24] HU, LI. "Animate anyone: Consistent and controllable image-to-video synthesis for character animation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 8153–8163 3, 6–8.

[HZ10] HORE, ALAIN and ZIOU, DJEMEL. "Image quality metrics: PSNR vs. SSIM". *2010 20th International Conference on Pattern Recognition*. IEEE. 2010, 2366–2369 6.

[JJH*26] JAVANMARDI, ALIREZA, JAISWAL, PRAGATI, HABTEGEBRIAL, TEWODROS AMBERBIR, et al. "TalkingPose: Efficient Face and Gesture Animation with Feedback-guided Diffusion Model". *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2026, 3098–3108 6, 7.

[JPS24] JAVANMARDI, ALIREZA, PAGANI, ALAIN, and STRICKER, DIDIER. "G3FA: Geometry-guided GAN for Face Animation". *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA, 2024 3.

[JSZ*25] JUNKAWITSCH, HENDRIK, SUN, GUOXING, ZHU, HEMING, et al. "EVA: Expressive Virtual Avatars from Multi-view Videos." (2025), 1–11 2.

[KB17] KINGMA, DIEDERIK P. and BA, JIMMY. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG] 6.

[KBM*24] KHIRODKAR, RAWAL, BAGAUTDINOV, TIMUR, MARTINEZ, JULIETA, et al. "Sapiens: Foundation for human vision models". *European Conference on Computer Vision*. Springer. 2024, 206–228 6.

[KGN24] KIRSCHSTEIN, TOBIAS, GIEBENHAIN, SIMON, and NIESSNER, MATTHIAS. "Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 5481–5492 2.

[KKLD23] KERBL, BERNHARD, KOPANAS, GEORGIOS, LEIMKÜHLER, THOMAS, and DRETTAKIS, GEORGE. "3D Gaussian Splatting for Real-Time Radiance Field Rendering". *ACM Transactions on Graphics* 42.4 (July 2023). URL: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/ 2, 5.

[KRS*25] KIRSCHSTEIN, TOBIAS, ROMERO, JAVIER, SEVASTOPOLSKY, ARTEM, et al. "Avat3r: Large Animatable Gaussian Reconstruction Model for High-fidelity 3D Head Avatars". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2025, 12089–12100 3.

[LBB*17] LI, TIANYE, BOLKART, TIMO, BLACK, MICHAEL. J., et al. "Learning a model of facial shape and expression from 4D scans". *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36.6 (2017), 194:1–194:17. URL: https://doi.org/10.1145/3130800.3130813 2, 6.

[LMR*15] LOPER, MATTHEW, MAHMOOD, NAUREEN, ROMERO, JAVIER, et al. "SMPL: A Skinned Multi-Person Linear Model". *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34.6 (Oct. 2015), 248:1–248:16 2.

[LTN*19] LUGARESI, CAMILLO, TANG, JIUQIANG, NASH, HADON, et al. "MediaPipe: A Framework for Perceiving and Processing Reality". *Proceedings of the Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*. 2019 6.

[LWP*24] LEI, JIAHUI, WANG, YUFU, PAVLAKOS, GEORGIOS, et al. "Gart: Gaussian articulated template models". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, 19876–19887 6.

[MLW*24] MA, YUE, LIU, HONGYU, WANG, HONGFA, et al. "Follow-Your-Emoji: Fine-Controllable and Expressive Freestyle Portrait Animation". *arXiv preprint arXiv:2406.01900* (2024) 2.

[MSS24] MOON, GYEONGSIK, SHIRATORI, TAKAAKI, and SAITO, SHUNSUKE. "Expressive Whole-Body 3D Gaussian Avatar". *ECCV*. 2024 2.

[MST*20] MILDENHALL, BEN, SRINIVASAN, PRATUL P., TANCIK, MATTHEW, et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". *ECCV*. 2020 2.

[PCG*19] PAVLAKOS, GEORGIOS, CHOUTAS, VASILEIOS, GHORBANI, NIMA, et al. "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image". *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, 10975–10985 4, 6.

[PSH*21] PARK, KEUNHONG, SINHA, UTKARSH, HEDMAN, PETER, et al. "HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields". *ACM Trans. Graph.* 40.6 (Dec. 2021) 2.

[PSR*24] PAVLAKOS, GEORGIOS, SHAN, DANDAN, RADOSAVOVIC, ILIJA, et al. "Reconstructing Hands in 3D with Transformers". *CVPR*. 2024 4, 6.

[PZK*24] PANG, HAOKAI, ZHU, HEMING, KORTYLEWSKI, ADAM, et al. "ASH: Animatable Gaussian Splats for Efficient and Photoreal Human Rendering". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, 1165–1175 2.

[QGL*25] QIU, LINGTENG, GU, XIAODONG, LI, PEIHAO, et al. "LHM: Large Animatable Human Reconstruction Model from a Single Image in Seconds". (2025) 2, 6.

[QKS*24] QIAN, SHENHAN, KIRSCHSTEIN, TOBIAS, SCHONEVELD, LIAM, et al. "Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 20299–20309 2, 4–6.

[RTB17] ROMERO, JAVIER, TZIONAS, DIMITRIOS, and BLACK, MICHAEL J. "Embodied Hands: Modeling and Capturing Hands and Bodies Together". *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*. 245:1–245:17 36.6 (Nov. 2017) 4, 6.

[SLT*19] SIAROHIN, ALIAKSANDR, LATHUILIÈRE, STÉPHANE, TULYAKOV, SERGEY, et al. "First order motion model for image animation". *Advances in Neural Information Processing Systems* 32 (2019) 3.

[SWR*21] SIAROHIN, ALIAKSANDR, WOODFORD, OLIVER J, REN, JIAN, et al. "Motion representations for articulated animation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, 13653–13662 2, 3.

[TRM*25] TEOTIA, KARTIK, RHODIN, HELGE, MENDIRATTA, MOHIT, et al. "Audio Driven Universal Gaussian Head Avatars". *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 2025, 1–12 2.

[TXH*25] TU, SHUYUAN, XING, ZHEN, HAN, XINTONG, et al. "Stableanimator: High-quality identity-preserving human image animation". *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, 21096–21106 2, 3.

[TZS*16] THIES, JUSTUS, ZOLLHOFER, MICHAEL, STAMMINGER, MARC, et al. "Face2Face: Real-Time Face Capture and Reenactment of RGB Videos". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. June 2016 2.

[TZTL25] TAUBNER, FELIX, ZHANG, RUIHANG, TULI, MATHIEU, and LINDELL, DAVID B. "CAP4D: Creating Animatable 4D Portrait Avatars with Morphable Multi-View Diffusion Models". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2025, 5318–5330 2, 3.

[WBSS04] WANG, ZHOU, BOVIK, ALAN C, SHEIKH, HAMID R, and SIMONCELLI, EERO P. "Image quality assessment: from error visibility to structural similarity". *IEEE Transactions on Image Processing* 13.4 (2004), 600–612 6.

[WLL*24] WANG, TAN, LI, LINJIE, LIN, KEVIN, et al. "Disco: Disentangled control for realistic human dance generation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 9326–9336 6.

[WML21] WANG, TING-CHUN, MALLYA, ARUN, and LIU, MING-YU. "One-shot free-view neural talking-head synthesis for video conferencing". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, 10039–10049 2, 3.

[WYBD22] WANG, YAOHUI, YANG, DI, BREMOND, FRANCOIS, and DANTCHEVA, ANTITZA. "Latent Image Animator: Learning to Animate Images via Latent Space Navigation". *International Conference on Learning Representations*. 2022 2, 3.

[WYBD24] WANG, YAOHUI, YANG, DI, BREMOND, FRANCOIS, and DANTCHEVA, ANTITZA. "LIA: Latent Image Animator". *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024), 1–16 3.

[XCL*24] XU, YUELANG, CHEN, BENWANG, LI, ZHE, et al. "Gaussian Head Avatar: Ultra High-fidelity Head Avatar via Dynamic Gaussians". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024 2, 6.

[XZL*24] XU, ZHONGCONG, ZHANG, JIANFENG, LIEW, JUN HAO, et al. "Magicanimate: Temporally consistent human image animation using diffusion model". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 1481–1490 3, 6–8.

[ZBT22] ZIELONKA, WOJCIECH, BOLKART, TIMO, and THIES, JUSTUS. "Towards metrical reconstruction of human faces". *European conference on computer vision*. Springer. 2022, 250–269 4, 6.

[ZCD*24] ZHU, SHENHAO, CHEN, JUNMING LEO, DAI, ZUOZHUO, et al. "Champ: Controllable and consistent human image animation with 3d parametric guidance". *arXiv preprint arXiv:2403.14781* (2024) 3, 6–8.

[ZGF*24] ZHENG, PENG, GAO, DEHONG, FAN, DENG-PING, et al. "Bilateral Reference for High-Resolution Dichotomous Image Segmentation". *CAAI Artificial Intelligence Research* 3 (2024), 9150038 6.

[ZGL*25] ZIELONKA, WOJCIECH, GARBIN, STEPHAN J., LATTAS, ALEXANDROS, et al. "Synthetic Prior for Few-Shot Drivable Head Avatar Inversion". *CVPR*. June 2025 2.

[ZGW*24] ZHANG, YUANG, GU, JIAXI, WANG, LI-WEN, et al. "MimicMotion: High-Quality Human Motion Video Generation with Confidence-aware Pose Guidance". *arXiv preprint arXiv:2406.19680* (2024) 3.

[ZLL*25] ZHANG, DONGBIN, LIU, YUNFEI, LIN, LIJIAN, et al. "GUAVA: Generalizable Upper Body 3D Gaussian Avatar". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2025, 14205–14217 2, 6–8.