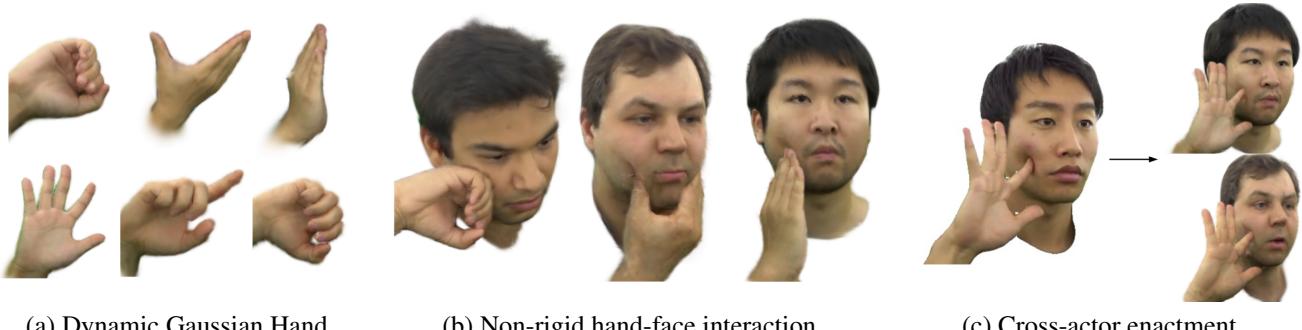


InteractAvatar: Modeling Hand-Face Interaction in Photorealistic Avatars with Deformable Gaussians

Kefan Chen^{1,2}Sreyas Mohan^{*2}Justin Theiss^{*2}Sergiu Oprea^{*2}Srinath Sridhar¹Aayush Prakash²¹Brown University²Meta Reality Labs

(a) Dynamic Gaussian Hand

(b) Non-rigid hand-face interaction.

(c) Cross-actor enactment.

Figure 1. We propose **InteractAvatar** which enables (a) **Dynamic Gaussian Hand**. Our novel representation anchors 3D Gaussian kernels to a hand template mesh and a learnable neural network allowing for pose-dependent articulation, self-cast shadows, and high-fidelity appearance modeling. (b) **Non-Rigid Hand-Face Interaction**. We introduce a learnable interaction module that refines hand-induced deformations and shading effects on the face, ensuring realistic skin contact dynamics. (c) **Cross-Actor Enactment**. We can transfer hand and face motions across different subjects, demonstrating its generalization capability to unseen identities and gestures.

Abstract

With increasing interest in digital avatars coupled with the importance of expressions and gestures in communication, modeling natural avatar behavior remains an important challenge in many industries such as teleconferencing, gaming, and AR/VR. Human hands are the primary tool for interacting with the environment and essential for realistic human behavior modeling, yet existing 3D hand and head avatar models often overlook the crucial aspect of hand-body interactions, such as between hand and face. We present **InteractAvatar**, the first model to faithfully capture the photorealistic appearance of dynamic hand and non-rigid hand-face interactions. Our novel **Dynamic Gaussian Hand** model, combining template model and 3D Gaussian Splatting as well as a dynamic refinement module, captures pose-dependent change, e.g. the fine wrinkles and complex shadows that occur during articulation. Importantly, our hand-face interaction module models the subtle geometry and appearance dynamics that underlie common gestures. Through experiments of novel view synthesis, self-reenactment, and cross-identity reenactment, we demon-

strate that **InteractAvatar** can reconstruct hand and hand-face interactions from monocular or multi-view videos with high-fidelity details and be animated with novel poses.

1. Introduction

Human communication transcends mere words with expressions, gestures, and subtle physical interactions conveying instantly recognizable emotions, intentions, and empathy. Studies show that these nonverbal cues play a crucial role in interaction [3], and failure to replicate them in digital avatars can disrupt comprehension and diminish immersion [2, 24]. Yet, faithfully capturing non-verbal communication modes in digital avatars remains a significant challenge. In fields like teleconferencing [17, 41, 42], VR/AR [33, 42, 47], gaming [5, 36], and virtual social worlds [42], there is a growing demand for avatars capable of dynamic, realistic interactions to enhance immersion and improve the quality of virtual human representation.

Although several methods exist to reconstruct photorealistic human avatars [1, 14, 19, 29, 50, 52, 53], they largely ignore hand-face interactions - an essential aspect of natural

human behavior. Studies show that people touch their face up to 600 [22] or 800 [45] times a day. Furthermore, simple, rigid interactions appear unnatural (see Figures 4, 5), underscoring the need for models that capture the subtleties of hand-face interactions.

Entertainment industries have traditionally relied on customized technologies like LightStage [7, 11, 25, 26, 40] or physics-based simulations [31, 48], such as finite element methods (FEM) [6, 15, 16] and position-based dynamics (PBD) [8, 9, 39, 46], to model the complex dynamics of skin deformation and contact between hands and face. However, these methods often require extremely specialized hardware for data capture, extensive manual parameter tuning, are computationally expensive, and often struggle to generalize across new poses or expressions. Data-driven approaches, including Neural Radiance Fields [27] and 3D Gaussian Splatting [20], are now state-of-the-art in modeling face and hand appearance [1, 14, 19, 29, 50, 52, 53], but do not yet handle hand-face interactions.

We introduce **InteractAvatar**, the first method to leverage 3D Gaussian Splatting for modeling **non-rigid hand-face deformations and interactions**. Our approach builds upon GaussianAvatar [35] by incorporating a hybrid mesh-Gaussian representation for both hands and faces, enabling high-fidelity animation and novel view synthesis. Our key innovations (see Figure 1) include:

- **Dynamic Gaussian Hands.** Capturing hand appearance and dynamics is challenging due to the complex geometry, varied textures, and intricate movements driven by numerous bones, joints, and muscles [44, 49]. We propose a *novel hand representation that anchors Gaussian kernels to a template mesh* and learns to *dynamically adjust geometry* (position, scale, rotation) and *appearance* (color, opacity). This allows accurate modeling of articulation-dependent effects like self-cast shadows and wrinkles, improving generalization to unseen poses.
- **Photorealistic Hand-Face Interaction Module.** While previous works have studied mesh recovery from hand-face interaction, our avatar model captures *fine-grained changes in facial geometry and appearance caused by hand contact*. This enables lifelike rendering of shadows, skin deformations, and pose-dependent visual effects.
- **New state-of-the-art.** Through extensive experiments on novel view synthesis, self-reenactment, and cross-identity reenactment, we demonstrate *qualitatively* and *quantitatively* that InteractAvatar achieves superior realism in modeling hand-face interactions compared to prior methods, *paving the way for more immersive digital experiences in AR/VR, gaming, and telepresence*.

2. Relevant Work

Photorealistic Avatars. Methods like Neural Radiance Fields (NeRF) [19] and 3D Gaussian Splatting [20], which

capture high-quality renderings of static scenes have been used to create photorealistic avatars [1, 14, 19, 29, 50, 52, 53]. Similar to these models, InteractAvatar builds on 3D Gaussians Splatting enabling high-fidelity reconstruction.

Neural Hand Rendering. Precise hand modeling is essential for realistic digital avatars, yet prior methods have key limitations. MANUS [32] uses articulated 3D Gaussian kernels for accurate hand-object contact but requires extensive multi-camera data and complex capture setups. Further, MANUS only focuses on interactions with rigid objects, unlike our task which requires much more intricate modeling. LiveHand [30], though capable of real-time, photorealistic rendering, adopts a NeRF-based implicit representation and does not efficiently model interactions between entities. LISA [10] offers a versatile model that combines shape, appearance, and animation, yet it lacks high-fidelity interaction modeling. In contrast, InteractAvatar’s *dynamic Gaussian hand* (a) attaches Gaussian kernel to template mesh obtained through MANO [38], and use (b) learnable modules that allow pose dependent deformations and changes, allows us to efficiently reason contact and deformation while generalizing to unseen poses during animation.

Hybrid Avatar Representations. Hybrid avatar representations leverage both mesh-based geometry and implicit models for realistic, adaptable avatars. SCARF [12] and DELTA [13] separate body structure from features like hair and clothing, blending meshes with neural fields to control each component individually. GaussianAvatar [?] and SplattingAvatar [43] embed Gaussian kernels on the FLAME [23] mesh, allowing them to deform smoothly with pose changes and capture high-detail appearances without predefined skinning weights. InteractAvatar adopts this approach. However, unlike static-feature approaches like Ref. [4, 18], which attach features to fixed mesh vertices, our model enable dynamic, pose-responsive deformations achieving better geometry and appearance.

Hand-Face Interaction. Reconstructing 3D hand-face interactions from images is challenging due to self-occlusions, diverse spatial relationships between hands and face, complex deformations, and the ambiguity of the single-view setting. DICE [49] estimates the poses of hands and faces, contacts, and deformations simultaneously using a Transformer-based architecture. DECAF [44], introduces a global fitting optimization guided by contact and deformation estimation networks trained on studio-collected data with 3D annotations. InteractAvatar uses a DECAF-like algorithm to resolve hand-face collisions and obtain coarse per-vertex 3D offsets, but refines these using a novel learnable module for precise facial geometry and appearance.

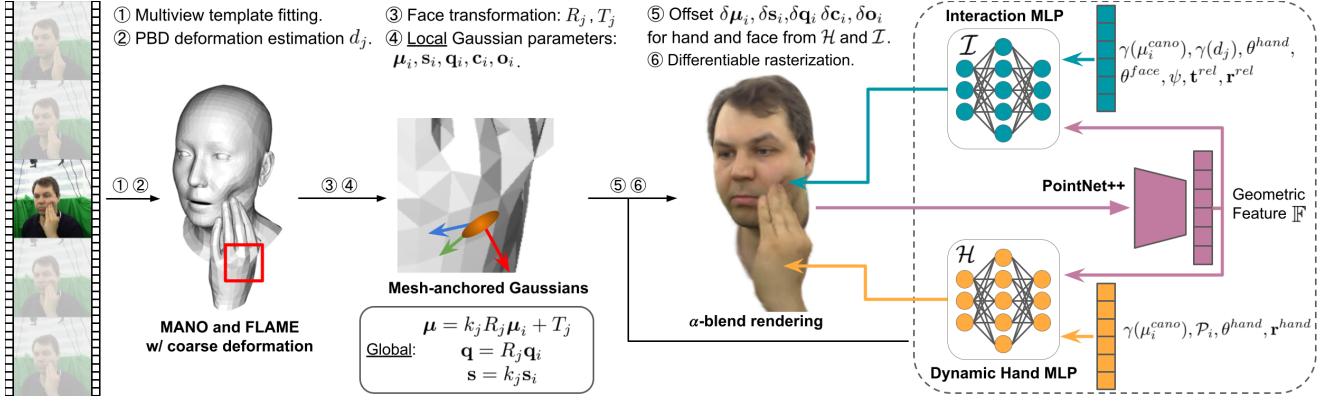


Figure 2. **Overview of InteractAvatar.** Our method combines mesh-based geometry (FLAME, MANO) with 3D Gaussian Splatting for realistic hand-face interactions. The dynamic hand appearance module refines pose-dependent deformations, wrinkles, and shadows, while the Hand-Face Interaction module enhances contact-aware geometry and shading adjustments. This enables high-fidelity animation with lifelike interactions and appearance changes.

3. Method

3.1. Preliminary

3D Gaussian Splatting [20] adopts explicit anisotropic Gaussians to reconstruct static scenes or objects from multi-view images and known camera parameters. Each Gaussian is defined by the position μ and a covariance matrix Σ representing shape:

$$G(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) \quad (1)$$

We can use a volume rendering formula similar to NeRF [27] to α -blend the N Gaussians and compute the color C of a pixel:

$$C(\mathbf{x}) = \sum_{i=1}^N \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad \alpha_i = o_i G'_i(\mathbf{x}) \quad (2)$$

where c_i is the color of each Gaussian point parameterized by spherical harmonics. The blending weight α is evaluated based on the opacity parameter o of the projected 2D Gaussians G' sorted by depth order.

3.2. InteractAvatar

We propose a novel approach to modeling hand-face interactions using an explicit hybrid mesh-Gaussian representation with dynamic deformation and appearance modules. Our method consists of: (1) a hybrid mesh-Gaussian avatar model for face and hand; (2) a dynamic hand appearance module that captures pose-dependent variations such as self-cast shadows and wrinkles during articulation; and (3) a hand-face interaction module that captures non-rigid deformation between these key body parts, as well as any deformation-induced changes in appearance.

Our hybrid representation combines the strengths of both the mesh and Gaussian models. The mesh provides a coarse estimation of deformation, and the Gaussians on the surface provide photorealistic appearance and changes in fine details. Specifically, we use FLAME [23] and MANO [38] as template mesh models that capture the shape, pose, and expression of the target human subject. We then build a layer of Gaussians on the mesh and bind the Gaussians to each mesh facet so that as the mesh moves and articulates, the Gaussians consistently follow, making it easy to animate and capture highly dynamic human subjects.

Our model is trained on multi-view videos of a subject performing hand-face interactions. At each time step, we get the FLAME and MANO mesh parameters for face and hand from multi-view template fitting [44], which requires multi-view observations and known camera calibration. We train the Gaussian avatar to reconstruct both hand and face, and use MLPs to model pose-dependent dynamic effects for the hand and interaction-induced deformation between hand and face.

Hybrid Mesh-Gaussian Face Avatar. Our face Gaussians are combined with the FLAME mesh that parameterizes the facial mesh via the pose θ^{face} , shape β^{face} , expression ψ , and global translation \mathbf{t}^{face} and rotation \mathbf{r}^{face} .

Following the definition in [35], for each mesh facet, we build a local coordinate frame based on the scale and orientation of the mesh. Each 3D Gaussian is bound to a mesh facet by assigning a face index. All spatial properties of the bound Gaussians are defined in the local triangle mesh and can freely move during optimization.

Globally, the Gaussians can be animated and articulated along with the mesh as the local mesh coordinate basis rotates and translates. We construct the local coordinate sys-

tem by setting the origin at the center of the triangle and form the orthogonal basis from the triangle normal and edge to obtain the rotation matrix R , which transforms the local Gaussians to the global world frame. We then define a scaling factor k proportional to the size of the triangle so that the local position and scaling of Gaussians are defined relative to the scale of their parent triangle. This also ensures that during optimization 3D Gaussians have updates that are proportional to their parent triangle size.

Specifically, for each Gaussian (indexed by i) on its parent triangle (indexed by j), we define the global transformation of the Gaussian as:

$$\boldsymbol{\mu} = k_j R_j \boldsymbol{\mu}_i + T_j \quad (3)$$

$$\mathbf{s} = k_j \mathbf{s}_i \quad (4)$$

$$\mathbf{q} = R_j \mathbf{q}_i \quad (5)$$

During optimization, the number of Gaussians needs to be adjusted to correctly fit the topology, so we adopt adaptive density control [20, 35] to populate or prune the Gaussians based on the view-space positional gradient and opacity. To preserve the binding to the correct local mesh coordinates as we clone or split a Gaussian bound to a mesh face, we assign the same parent mesh index to the new Gaussian so that all new Gaussians are sampled close to the old ones for a smooth optimization process.

Hybrid Mesh-Gaussian Hand Avatar. Our hand Gaussians are similarly combined with the MANO mesh that parameterizes the hand topology via the estimated pose θ^{hand} , shape β^{hand} , and global translation \mathbf{t}^{hand} and rotation \mathbf{r}^{hand} . Note that, to the best of our knowledge, this is the first avatar work to obtain a hybrid representation of hands by anchoring Gaussian kernels to mesh. Previous work like MANUS [32] anchors Gaussian kernels to a skeleton and learns inverse skinning weights but not to the mesh.

As described above for our face Gaussians, we follow the same procedure to build a layer of Gaussians to represent the hand on the MANO mesh and use the transformations outlined in Eqs. 3-5 to obtain the global position, scale, and rotation of the Gaussians when animating the hand.

Dynamic Hand Appearance Refinement. Compared to the face, hands present a unique challenge due to complex articulations and visual dynamics, such as self-cast shadows, wrinkles, and intricate skin deformations. Static Gaussian parameters are therefore not sufficient to capture high-fidelity details for hands. To address this gap, we model geometry and appearance dynamics using a set of MLPs to estimate offset on each Gaussian parameter:

$$\delta\boldsymbol{\mu}_i, \delta\mathbf{s}_i, \delta\mathbf{q}_i = \mathcal{H}_{geo}(\gamma(\boldsymbol{\mu}_i^{cano}), \theta^{hand}) \quad (6)$$

$$\delta\mathbf{c}_i, \delta\mathbf{o}_i = \mathcal{H}_{app}(\gamma(\boldsymbol{\mu}_i^{cano}), \mathcal{P}_i, \mathbb{F}, \theta^{hand}, \mathbf{r}^{hand}, \mathbf{r}^{rel}, \mathbf{t}^{rel}) \quad (7)$$

where $\gamma(\boldsymbol{\mu}_i^{cano})$ is the positional encoding of the canonical position of the Gaussian (set by the first frame of the training sequence), $\mathcal{P}_i \in \mathbb{R}^{64}$ is the per point feature associated with each Gaussian that is optimized together with the MLPs, \mathbb{F} encodes the geometric features which is detailed in the next section, while \mathbf{r}^{rel} and \mathbf{t}^{rel} are the relative orientation and translation between the hand and face which models interaction-induced effects on hand.

The geometry MLP \mathcal{H}_{geo} captures the intricate pose-dependent shape dynamics represented by the relative change in position ($\delta\boldsymbol{\mu}_i$), scale ($\delta\mathbf{s}_i$), and rotation ($\delta\mathbf{q}_i$) of each Gaussian, which are added to the corresponding variables in Eqs. 3-5.

The appearance MLP \mathcal{H}_{app} uses an auto-decoder to capture the appearance change (*i.e.* relative change in color $\delta\mathbf{c}_i$ and opacity $\delta\mathbf{o}_i$) dependent on the hand pose θ^{hand} and orientation \mathbf{r}^{hand} . We further incorporate a point feature embedding $\mathcal{P}_i \in \mathbb{R}^{64}$ as an additional parameter for each Gaussian that is jointly optimized with the MLP. We use \mathcal{H}_{app} to update color and opacity as $\mathbf{c}_i = \mathbf{c}_i + \delta\mathbf{c}_i$ and $\mathbf{o}_i = \mathbf{o}_i + \delta\mathbf{o}_i$, respectively.

The relative changes in each parameter are then added to obtain the resultant Gaussian parameters:

$$\boldsymbol{\mu} = k_j R_j (\boldsymbol{\mu}_i + \delta\boldsymbol{\mu}_i) + T_j \quad (8)$$

$$\mathbf{s} = k_j (\mathbf{s}_i + \delta\mathbf{s}_i) \quad (9)$$

$$\mathbf{q} = R_j (\mathbf{q}_i + \delta\mathbf{q}_i) \quad (10)$$

$$\mathbf{c}_i = \mathbf{c}_i + \delta\mathbf{c}_i \quad (11)$$

$$\mathbf{o}_i = \mathbf{o}_i + \delta\mathbf{o}_i \quad (12)$$

Hand-Face Interaction. To simulate the deformation changes due to hand and face interactions, we can leverage a physical simulation method like position-based dynamics (PBD) [28], which avoids overfitting and generalization challenges when learning physical deformations from data. We follow DECAF [44] to resolve hand-face collisions and obtain per-vertex 3D offsets of the facial geometry of interacting hand-face deformation. To obtain more anatomically realistic deformation, DECAF [44] computes the stiffness value for each face vertex based on skin-skull distance. As the facial mesh deforms, the local Gaussians also deform along with the mesh. However, the coarsely estimated deformation may present a gap between the predicted and actual deformations. To address this deformation gap as well as appearance changes (e.g., shadows), we propose a set of interaction MLPs:

$$(\delta\boldsymbol{\mu}_i, \delta\mathbf{s}_i, \delta\mathbf{q}_i, \delta\mathbf{c}_i, \delta\mathbf{o}_i) = \mathcal{I}(\gamma(\mu_i^{cano}), \gamma(d_j), \mathbb{F}, \theta^{hand}, \theta^{face}, \psi, \mathbf{t}^{rel}, \mathbf{r}^{rel}) \quad (13)$$

where, similar to the dynamic Gaussian hand, the deformation dynamics are modeled as relative change in position, scale, and rotation for each Gaussian and appearance dynamics are modeled as the relative change in color and opacity. The hand-face interaction MLPs \mathcal{I} take as input the positional encoding of the canonical position $\gamma(\mu_i^{cano})$ and deformation offset from the parent mesh face $\gamma(d_j)$, the global geometric feature \mathbb{F} , the pose of the hand θ^{hand} and face θ^{face} , facial expression ψ (estimated from FLAME), as well as the relative translation \mathbf{t}^{rel} and rotation \mathbf{r}^{rel} between the hand and face. To compute the deformation offset d_j , we take the average offset of the parent mesh face vertices $d_j = \frac{1}{3} \sum (V_1 + V_2 + V_3)$.

In order to capture the local and global relationships among hand and face Gaussians, we use PointNet++ [34] to extract the geometric feature $\mathbb{F} \in \mathbb{R}^{1024}$. However, it is computationally infeasible to consider hundreds of thousands of Gaussians in this manner, so we instead propose an efficient sampling heuristic. First, we consider only the Gaussians located at non-rigid facial regions where deformation is possible. Next, we sample a single Gaussian from each mesh facet proportional to $\mathbf{o}_i \|\mathbf{s}_i\|_2$, since the Gaussian with highest opacity and largest scale contributes most to the local geometry.

Since some regions of the face are more non-rigid than others (*e.g.* cheek), we can focus updates within these regions to improve computational efficiency by selecting only the mesh faces that have a deformation greater than a minimal threshold. Intuitively, larger deformations occur in regions closest to the contact point between the hand and face. We therefore apply weights w_i to the predicted offsets inversely proportional to the distance between the Gaussian and nearest hand vertex, as shown in the following equations.

$$w_i = \frac{1}{2} (\cos(\pi \frac{\min_j \|\boldsymbol{\mu}_i - \mathbf{V}_j\|_2}{d_{max}}) + 1) \quad (14)$$

$$\mathbf{x}_i = \mathbf{x}_i + w_i \delta \mathbf{x}_i, \quad \mathbf{x}_i = \{\boldsymbol{\mu}_i, \mathbf{s}_i, \mathbf{q}_i, \mathbf{c}_i, \mathbf{o}_i\} \quad (15)$$

where $d_{max} = 0.05$ is the maximum distance and anything exceeds this limit has $w_i = 0$.

3.3. Training and Regularization

Our method follows a two-stage training paradigm with an initial warm-up phase to train the static Gaussian parameters followed by training of the hand and interaction MLPs

\mathcal{H} and \mathcal{I} . In the first stage, we set the first frame in the training sequence as the canonical frame and initialize N Gaussians at the center of each face on the hand and face meshes. We initialize the rest of the Gaussian parameters as done in [20]. We supervise the rendered images with \mathcal{L}_1 and SSIM losses:

$$\mathcal{L} = (1 - \lambda) \mathcal{L}_1 + \lambda \mathcal{L}_{D-SSIM} \quad (16)$$

During the second stage of training, we initialize the weights of the final layers of the hand and interaction MLPs with zeros in order to fine-tune intricate details without negatively impacting the overall structure early in training.

When training the hand and interaction MLPs, we identify a few challenges that must be addressed. First, since hand-face interactions are relatively rare occurrences within video sequences, we oversample video frames where such interactions do occur to ensure sufficient training data. Second, when animating the Gaussian avatar to novel poses, we occasionally encounter thin and spiky Gaussians, which are exacerbated on the hand by significant articulations of the fingers. Intuitively, larger Gaussians are more likely to contribute to such visual artifacts because a small rotation could be greatly magnified by the scale. To mitigate this, we regularize the scaling to encourage smaller and more isotropic Gaussians:

$$\mathcal{L}_s = \|ReLU(\mathbf{s} + \delta \mathbf{s} - \epsilon_s)\|_2^2 \quad (17)$$

where ϵ_s is a minimal threshold for the scaling parameter to prevent the Gaussians from shrinking excessively.

Third, Gaussians can deviate from the coarse topology of the mesh in an unrealistic manner. To prevent such deviation from the parent mesh, we further add a position regularization loss:

$$\mathcal{L}_\mu = \|ReLU(\boldsymbol{\mu} + \delta \boldsymbol{\mu} - \epsilon_\mu)\|_2^2 \quad (18)$$

where ϵ_μ defines the area around the mesh that Gaussians are free to move about in order to adjust the discrepancies between the fitted mesh and visual observation. We only apply \mathcal{L}_s and \mathcal{L}_μ to visible Gaussians to preserve the structure of occluded parts as the hand articulates and interacts with the face.

Finally, since hands and hand-face interactions usually comprise a small area of the scene yet have rich visual details, we track the hand and face with bounding boxes and add a patch perceptual loss \mathcal{L}_{patch} [51] that targets the hand as well as the overlapping region of hand and face bounding boxes. Our overall training loss is defined as follows:

$$\mathcal{L} = (1 - \lambda) \mathcal{L}_1 + \lambda \mathcal{L}_{D-SSIM} + a \mathcal{L}_s + b \mathcal{L}_\mu + c \mathcal{L}_{patch} \quad (19)$$

4. Experiment

4.1. Setup

Implementation details. We follow [44] to obtain MANO and FLAME mesh parameters and coarse mesh deformation from multiview videos by fitting 2D projected keypoints given known camera calibrations and solving PBD optimization. We initialize 20 Gaussians per mesh face on the mesh, assume no view-dependent effects, and disable spherical harmonics. We use 4-layer and 6-layer MLPs with 256 hidden units and leaky ReLU activation, layer norm, and a skip connection from the input to the middle layer for the dynamic hand module \mathcal{H} and interaction module \mathcal{I} respectively. We designate an MLP for each Gaussian parameter. We first train the static Gaussian face and hand avatars for $100k$ steps as a warm-up before training the hand and hand-face interaction MLPs. We use the position embedding in [27] to encode Gaussian positions.

We use Adam [21] to optimize our model and use the same learning rate and exponential learning decay as in the original 3D Gaussian Splatting [20]. We set the learning rate for MLPs to $1e - 3$ and the learning rate for the point feature $2.5e - 3$. We set loss weights $\lambda = 0.2$, $a = 1.0$, $b = 0.01$, $c = 0.1$ and the threshold $\epsilon_s = 0.4$ and $\epsilon_\mu = 0.2$. We train InteractAvatar on multi-view videos using an NVIDIA RTX 3090 GPU.

Datasets and evaluation. We evaluate novel view reconstruction, self-enactment on novel sequences, and cross-actor enactment of hand-face interaction using the only public available 3D hand-face interaction dataset DECAF [44]. We take four subjects and use seven views to train while holding out one view to test novel views. We use the train and test split in DECAF for self-enactment and cross-actor enactment evaluation. We use SAM 2 [37] to segment the hand and face then crop and resize the images to 512×512 using the tracked hand and face bounding boxes. For quantitative metrics, we report PSNR, SSIM, and LPIPS.

Baselines We select recent state-of-the-art Gaussian face avatars like GaussianAvatar [35] and SplattingAvatar [43] as our baselines. They are both the latest hybrid Gaussian head avatars anchored on a template face or body mesh. We naively extend their representation to modeling hands driven by MANO hand template mesh in our setting.

4.2. Dynamic Gaussian Hand

In Figure 3 and bottom row in Figure 4, we demonstrate that our proposed hand module captures high-fidelity details of pose-dependent dynamics such as self-cast shadows and wrinkles. The baseline GaussianAvatar and SplattingAvatar only allow Gaussian geometry to morph based on the transformation underlying mesh they are bounded to. This

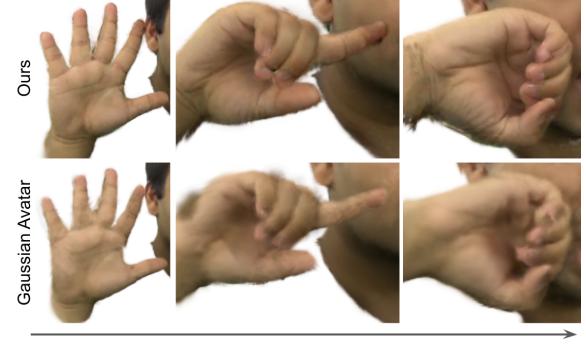


Figure 3. **Dynamic Gaussian Hand** adapts to pose, capturing self-cast shadows, wrinkles, and shading variations. The baseline methods struggle with static hand modeling, whereas our approach preserves fine-grained details across diverse hand poses.

		PSNR↑	SSIM↑	LPIPS↓
Novel view	SplattingAvat. [43]	25.90	0.9135	0.0584
	GaussianAvat. [35]	26.71	0.9324	0.0542
	Ours	29.85	0.9335	0.0338
Self-enact	SplattingAvat. [43]	26.18	0.9326	0.0546
	GaussianAvat. [35]	25.51	0.9393	0.0578
	Ours	28.17	0.9356	0.0396

Table 1. **InteractAvatar outperforms baselines quantitatively** on PSNR and LPIPS across novel view synthesis and self-enactment.

	PSNR↑	SSIM↑	LPIPS↓
Ours	29.85	0.9335	0.0338
w/o hand MLPs \mathcal{H}	26.97	0.9359	0.0574
w/o interaction MLPs \mathcal{I}	28.23	0.9389	0.0467

Table 2. **Ablation Study on Novel View Synthesis.** Our full model provides the best perceptual quality.

limits the expressiveness of the Gaussian geometry to capture highly articulated hands and assumes static appearance as color and opacity remain the same. Thanks to the dynamic hand MLPs \mathcal{H} , we can learn the change in geometry and appearance dependent on hand pose and orientation.

4.3. Hand-Face Interaction

We evaluate the quality of reconstruction on held-out novel views (Figure 4) and animation quality on self-reenactment (Figure 5) using hand and head poses in test sequence to drive the trained avatar. We show quantitative evaluation in Table 1. While PSNR and LPIPS are greatly improved, SSIM in our self-enactment is slightly worse than Gaussian Avatar, because our major improvements focus on sharp dynamic details on hand and interaction region but the MLPs may introduce some noises in the global structure of face and hand in test sequences due to generalization gap. Both Splatting Avatar and Gaussian Avatar show great quality in face reconstruction and animation. However, they struggle to reconstruct hands due to their complex articulation and dynamic appearance compared with facial

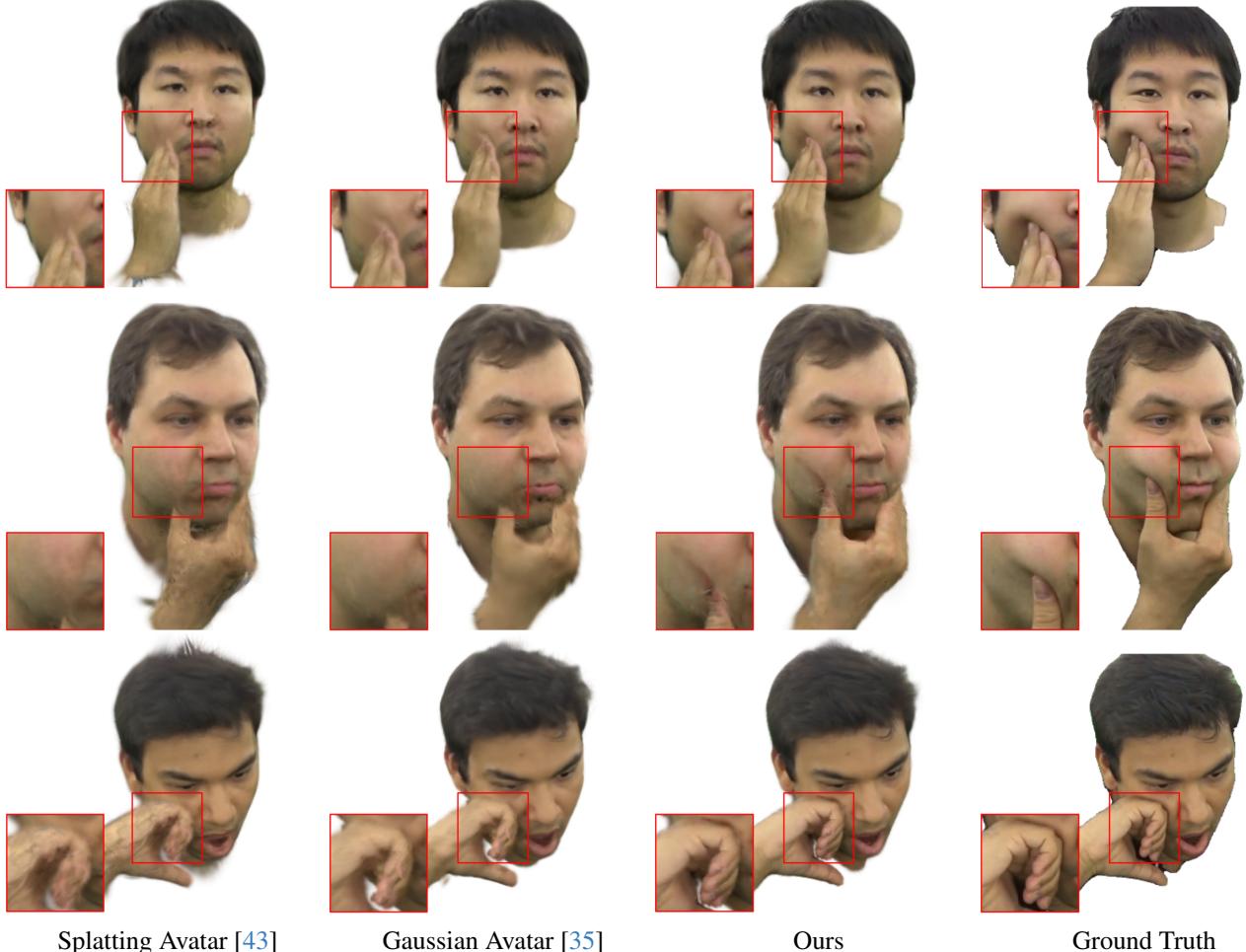


Figure 4. Qualitative Comparison of Hand-Face Interactions from Novel Views. Our method produces sharp, high-fidelity details on non-rigid facial deformations and dynamic hand appearances, outperforming baseline models like GaussianAvatar [35] and SplattingAvatar [43]. Features like shadowing, wrinkles, and natural hand-face deformations are accurately reconstructed.

expressions. Splatting Avatar allows Gaussians to walk between the mesh faces yet demonstrate worse quality than Gaussian Avatar which bounds the Gaussians to particular mesh faces. Gaussian Avatar shows over-smoothed appearance on hands as it doesn't model dynamic color and opacity as the geometry changes. Both baselines fail to capture the non-rigid deformation and appearance dynamics when hand interacts with face. Our method captures the complex dynamics of both hand and hand-face interaction by encoding pose and geometry with MLPs. In Figure 6, we demonstrate that our interaction module can be generalized to motions from different actors, showcasing applications in film, gaming, and VR. We did ablation studies on subject #4 and report results in Table 2 and Figure 7, demonstrating the effects of each design choice. Particularly, without our interaction module, naively training with PBD mesh deformation only produces artifacts in Gaussian geometry shown in the 2nd example of Figure 7.

4.4. Limitations

Despite some promising results, it is difficult to handle the vast complexity of hand-face interactions. While the dataset we use captures some common behaviors, it cannot exhaustively cover the countless possible configurations of hand-face dynamics, potentially limiting generalization to novel scenarios in the wild. Given the scarcity of specialized datasets for human self-interaction, we hope our research will bring greater attention to this important yet underexplored problem within the research community.

5. Conclusion

We address a valuable and challenging problem for entertainment and social media industries: hand-face interaction for human avatars. We present InteractAvatar, the first method to model dynamic hands and hand-face interactions. We do so by proposing a novel hybrid mesh-Gaussian representation of hands that models deformations and ap-

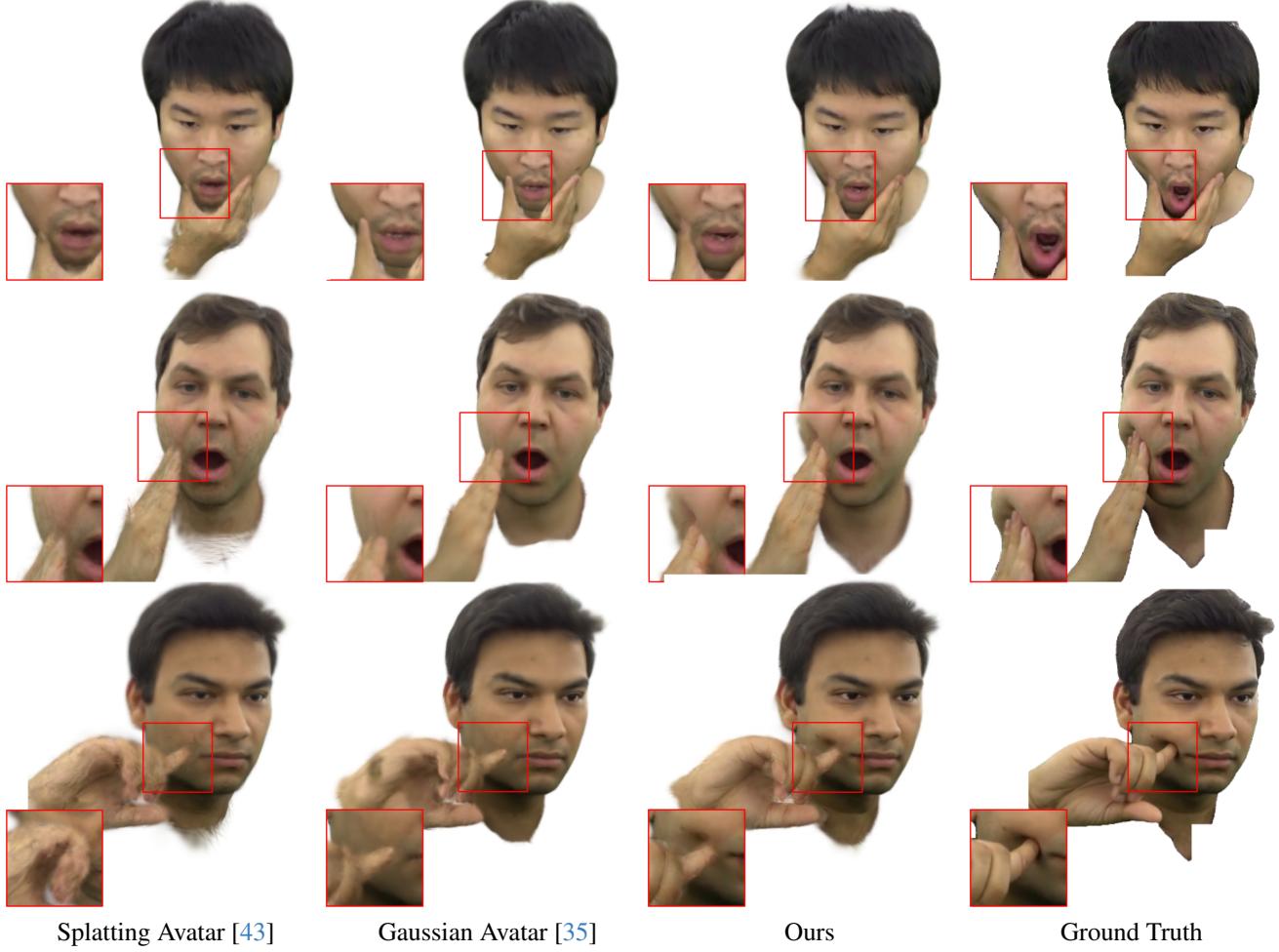


Figure 5. Self-Enactment with InteractAvatar. Our method accurately reconstructs natural hand-face interactions, preserving fine geometric details and appearance consistency in self-enactment tasks. Compared to baselines, InteractAvatar effectively models dynamic wrinkles, shadows, and subtle hand-induced deformations.

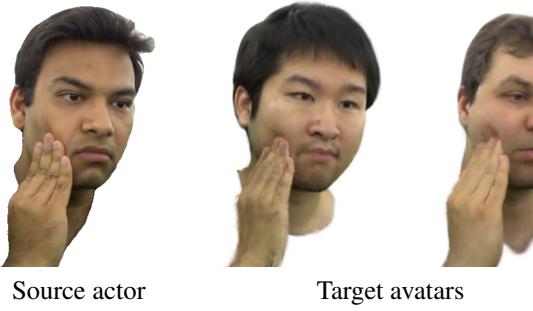


Figure 6. Cross-Actor Re-enactment. We use the pose θ^{face} , θ^{hand} and expression ψ parameters from the tracked FLAME and MANO templates of the source actor to drive the target avatars. Our InteractAvatar can generalize to novel motions and interaction-induced dynamics.

pearance dynamics during articulation. We further propose a novel hand-face interaction module that models geometry and appearance changes during hand-face interaction.

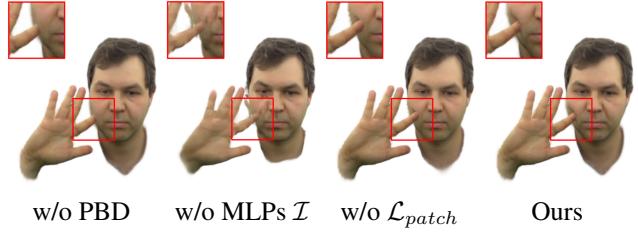


Figure 7. Ablation. Red highlights: (1) PBD provides coarse geometric deformation on mesh. (2) Naively training with PBD deformation without the interact MLPs \mathcal{I} to compensate the geometry would result in "flying out" Gaussians. (3) The interaction regions are small in the image space and provides weak training signal, we add a patch perceptual loss \mathcal{L}_{patch} to better capture the fine visual details.

Our technique can model realistic deformation, render complex details like shadows, generalize to unseen poses, and enable reenactment. Future work will focus on improving generalization for complex hand-face interactions.

References

- [1] Instant Volumetric Head Avatars, 2023. [1](#), [2](#)
- [2] Alex Adkins, Aline Normoyle, Lorraine Lin, Yu Sun, Yuting Ye, Massimiliano Di Luca, and Sophie Joerg. How important are detailed hand motions for communication for a virtual character through the lens of charades? *ACM Transactions on Graphics*, 42(3):1–16, 2023. [1](#)
- [3] Esma Nur Asalioğlu and Tilbe Göksun. The role of hand gestures in emotion communication: Do type and size of gestures matter? *Psychological Research*, 87(6):1880–1898, 2023. [1](#)
- [4] Ziqian Bai, Feitong Tan, Zeng Huang, Kripasindhu Sarkar, Danhang Tang, Di Qiu, Abhimitra Meka, Ruofei Du, Mingsong Dou, Sergio Orts-Escalano, Rohit Pandey, Ping Tan, Thabo Beeler, Sean Fanello, and Yinda Zhang. Learning personalized high quality volumetric head avatars from monocular rgb videos, 2023. [2](#)
- [5] Jaime Banks and Nicholas David Bowman. Emotion, anthropomorphism, realism, control: Validation of a merged metric for player–avatar interaction (pax). *Computers in Human Behavior*, 54:215–223, 2016. [1](#)
- [6] Sumit Basu, Nuria Oliver, and Alex Pentland. 3d modeling and tracking of human lip motions. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 337–343. IEEE, 1998. [2](#)
- [7] P Bebevec. The light stage: Photorealistically integrating real actors into virtual environments. *Svenska Föreningen för Grafisk Databehandling*, pages 17–22, 2002. [2](#)
- [8] Jan Bender, Matthias Müller, and Miles Macklin. Position-based simulation methods in computer graphics. In *Eurographics (tutorials)*, page 8, 2015. [2](#)
- [9] Nuttapong Chentanez, Miles Macklin, Matthias Müller, Stefan Jeschke, and Tae-Yong Kim. Cloth and skin deformation with a triangle mesh based convolutional neural network. In *Computer Graphics Forum*, pages 123–134. Wiley Online Library, 2020. [2](#)
- [10] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. Lisa: Learning implicit shape and appearance of hands. In *CVPR*, 2022. [2](#)
- [11] Paul Debevec, Andreas Wenger, Chris Tchou, Andrew Gardner, Jamie Waese, and Tim Hawkins. A lighting reproduction approach to live-action compositing. *ACM Transactions on Graphics (TOG)*, 21(3):547–556, 2002. [2](#)
- [12] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. [2](#)
- [13] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J. Black. Learning disentangled avatars with hybrid 3d representations. *arXiv*, 2023. [2](#)
- [14] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video.
- ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia), 41(6), 2022. [1](#), [2](#)
- [15] J-P Gourret, Nadia Magnenat Thalmann, and Daniel Thalmann. Simulation of object and human skin formations in a grasping task. In *Proceedings of the 16th annual conference on Computer graphics and interactive techniques*, pages 21–30, 1989. [2](#)
- [16] Brian Guenter. *A system for simulating human facial expression*. Springer, 1989. [2](#)
- [17] Simon NB Gunkel, Hans M Stokking, Martin J Prins, Nanda Van Der Stap, Frank B ter Haar, and Omar A Niamut. Virtual reality conferencing: Multi-user immersive vr experiences on the web. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 498–501, 2018. [1](#)
- [18] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans, 2023. [2](#)
- [19] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. [1](#), [2](#)
- [20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. [2](#), [3](#), [4](#), [5](#), [6](#)
- [21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. [6](#)
- [22] Yen Lee Angela Kwok, Jan Gralton, and Mary-Louise McLaws. Face touching: a frequent habit that has implications for hand hygiene. *American journal of infection control*, 43(2):112–114, 2015. [2](#)
- [23] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. [2](#), [3](#)
- [24] Pengcheng Luo and Michael Neff. A perceptual study of the relationship between posture and gesture for virtual characters. In *Motion in Games: 5th International Conference, MIG 2012, Rennes, France, November 15-17, 2012. Proceedings 5*, pages 254–265. Springer, 2012. [1](#)
- [25] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Maeta, Andrew Jewett, Simion Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matthew Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail,

- Melissa Schoeller, and Yaser Sheikh. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. In *Advances in Neural Information Processing Systems*, pages 83008–83023. Curran Associates, Inc., 2024. 2
- [26] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshatin, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Hidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024. 2
- [27] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3, 6
- [28] Matthias Müller, Bruno Heidelberger, Marcus Hennix, and J. W. Ratcliff. Position based dynamics. In *Journal of Visual Communication and Image Representation*, 2007. 4
- [29] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 1, 2
- [30] Akshay Mundra, Mallikarjun B R, Jiayi Wang, Marc Habermann, Christian Theobalt, and Mohamed Elgarib. Live-hand: Real-time and photorealistic neural hand rendering, 2023. 2
- [31] Jun-yong Noh and Ulrich Neumann. A survey of facial modeling and animation techniques. Technical report, USC Technical Report, 99–705, 1998. 2
- [32] Chandradeep Pokharia, Ishaan Nikhil Shah, Angela Xing, Zekun Li, Kefan Chen, Avinash Sharma, and Srinath Sridhar. Manus: Markerless grasp capture using articulated 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2197–2208, 2024. 2, 4
- [33] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfelix, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018. 1
- [34] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. 5
- [35] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 2, 3, 4, 6, 7, 8
- [36] Rabindra A Ratan and Michael Dawson. When mii is me: A psychophysiological examination of avatar self-relevance. *Communication Research*, 43(8):1065–1093, 2016. 1
- [37] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 6
- [38] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2, 3
- [39] Valentin Roussellet, Nadine Abu Rumman, Florian Canezin, Nicolas Mellado, Ladislav Kavan, and Loïc Barthe. Dynamic implicit muscles for character skinning. *Computers & Graphics*, 77:227–239, 2018. 2
- [40] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *CVPR*, 2024. 2
- [41] Mike Seymour, Kai Riener, and Judy Kay. Actors, avatars and agents: Potentials and implications of natural face technology for the creation of realistic visual presence. *Journal of the association for Information Systems*, 19(10):4, 2018. 1
- [42] Mike Seymour, Lingyao Yuan, Alan Dennis, and Kai Riener. Facing the artificial: Understanding affinity, trustworthiness, and preference for more realistic digital humans. 2020. 1
- [43] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 6, 7, 8
- [44] Soshi Shimada, Vladislav Golyanik, Patrick Pérez, and Christian Theobalt. Decaf: Monocular deformation capture for face and hand interactions. *ACM Transactions on Graphics (TOG)*, 42(6), 2023. 2, 3, 4, 6
- [45] Jente L Spille, Martin Grunwald, Sven Martin, and Stephanie M Mueller. Stop touching your face! a systematic review of triggers, characteristics, regulatory functions and neuro-physiology of facial self touch. *Neuroscience & Biobehavioral Reviews*, 128:102–116, 2021. 2
- [46] Mingyang Sun, Dongliang Kou, Ruisheng Yuan, Dingkang Yang, Peng Zhai, Xiao Zhao, Yang Jiang, Xiong Li, Jingchen Li, and Lihua Zhang. Physhand: A hand simulation model with physiological geometry, physical deformation, and ac-

- curate contact handling. *arXiv preprint arXiv:2409.05143*, 2024. [2](#)
- [47] Shih-En Wei, Jason Saragih, Tomas Simon, Adam W Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. Vr facial animation via multiview image translation. *ACM Transactions on Graphics (ToG)*, 38(4):1–16, 2019. [1](#)
- [48] Nkenge Wheatland, Yingying Wang, Huaguang Song, Michael Neff, Victor Zordan, and Sophie Jörg. State of the art in hand and finger modeling and animation. In *Computer Graphics Forum*, pages 735–760. Wiley Online Library, 2015. [2](#)
- [49] Qingxuan Wu, Zhiyang Dou, Sirui Xu, Soshi Shimada, Chen Wang, Zhengming Yu, Yuan Liu, Cheng Lin, Zeyu Cao, Taku Komura, Vladislav Golyanik, Christian Theobalt, Wenping Wang, and Lingjie Liu. Dice: End-to-end deformation capture of hand-face interactions from a single image, 2024. [2](#)
- [50] Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. Avatarmav: Fast 3d head avatar reconstruction using motion-aware neural voxels. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. [1](#), [2](#)
- [51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. [5](#)
- [52] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. [1](#), [2](#)
- [53] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [1](#), [2](#)