CaricatureGS: Exaggerating 3D Gaussian Splatting Faces With Gaussian Curvature

Eldad Matmon Amit Bracha Noam Rotstein Ron Kimmel

Technion – Israel Institute of Technology, Haifa, Israel

Figure 1. Photorealistic 3D caricature avatars produced by our method.

Abstract

A photorealistic and controllable 3D caricaturization framework for faces is introduced. We start with an intrinsic Gaussian curvature-based surface exaggeration technique, which, when coupled with texture, tends to produce over-smoothed renders. To address this, we resort to 3D Gaussian Splatting (3DGS), which has recently been shown to produce realistic free-viewpoint avatars. Given a multiview sequence, we extract a FLAME mesh, solve a curvature-weighted Poisson equation, and obtain its exaggerated form. However, directly deforming the Gaussians yields poor results, necessitating the synthesis of pseudo–ground-truth caricature images by warping each frame to its exaggerated 2D representation using local affine transformations. We then devise a training scheme that alternates real and synthesized supervision, enabling a single Gaussian collection to represent both natural and exaggerated avatars. This scheme improves fidelity, supports local edits, and allows continuous control over the intensity of the caricature. In order to achieve real-time deformations, an efficient interpolation between the original and exaggerated surfaces is introduced. We further analyze and show that it has a bounded deviation from closed-form solutions. In both quantitative and qualitative evaluations, our results outperform prior work, delivering photorealistic, geometry-controlled caricature avatars.

Project page: https://c4ricaturegs.github.io

1. Introduction

Face caricaturization refers to the action of exaggerating distinctive facial features while preserving identity. Despite its promise for lifelike, immersive avatars, producing such exaggerations in controllable, photorealistic 3D remains an open challenge. Successful mesh-based approaches are based on geometric deformations with curvature-based methods, such as the scale-aware Poisson framework [29]. When such deformed surfaces are rendered through traditional mesh-centric pipelines, such as texture mapping, the results often appear unnatural [29]. Recently, 3D Gaussian Splatting (3DGS) [16] has emerged as a potential multiview representation that provides state-of-the-art real-time photorealism by optimizing Gaussian primitives directly from a given set of images taken from various directions.

This raises the following question.

Can we combine curvature-based geometric fidelity with 3DGS to generate photorealistic caricatures?

To address this, we start with a multiview video of a subject and its extracted FLAME mesh [21]. From this, solving the weighted Poisson equation gives us the deformed caricature mesh. We rig Gaussians to the original undeformed surface and train them following a framework previously proposed for facial expressions [19]. Later, at inference, we deform the original mesh and its rigged Gaussians according to the caricature mesh, stretching, shearing, and rotating them. However, modeling these deformations as merely an additional expression, using Gaussians optimized only on

the input sequence, leads to low fidelity (see Fig. 5), revealing a domain gap in which caricatures lie outside the distribution of natural expression dynamics.

To bridge this gap and in the absence of real caricature training data, we synthesize pseudo–ground truth (GT*) by warping each input frame with *Local Affine Transformations* (LAT) induced by the correspondence from the original mesh to its curvature-exaggerated counterpart, producing photorealistic supervision (see Sec. 3.2). During training, we stochastically alternate between real views and GT*views so that a single Gaussian set jointly models both natural and caricatured deformations, allowing the Gaussians to benefit from real ground truth while adapting to GT*. To mitigate occlusion-related artifacts and protect fine structures (*e.g*. hair and mesh boundaries), we apply a spatial mask that freezes the affected Gaussians during GT*steps (Fig. 7). These Gaussians are updated only from real frames, allowing a consistent appearance to accumulate in their attributes.

Although trained only on the two sets of views, the optimized model offers additional flexibility and control at inference. First, it generalizes across a continuous range of caricature intensities, with the exaggeration level controlled by an efficient linear interpolation as an approximation of the solution to the weighted Poisson equation, a property that we demonstrate both theoretically and empirically. Moreover, this representation is robust to both global and local deformations, enabling controlled localized edits, such as exaggerating the nose size, while leaving unrelated regions unchanged.

The new 3DGS animatable representation is the first, to our knowledge, to enable photorealistic caricature rendering while faithfully retaining identity under caricature deformations. We compare it to the current state-of-the-art dynamic facial reconstruction model [19], which consistently achieves higher scores and qualitative results in terms of image fidelity, structural consistency, and identity preservation metrics.

Our contributions include,

- A novel 3DGS training scheme that uses GT*generated with local Affine transformations that represent real and caricature avatars.
- Curvature-weighted deformation with rigged 3DGS for identity-preserving photorealistic caricatures.
- Real-time avatars supporting variable exaggeration levels and fine-grained local control of facial features.

2. Related Work

2.1. Representation for 3D Head Avatars

Neural implicit representations have become a dominant approach for high-fidelity 3D head avatars, enabling photorealistic view synthesis from sparse multiview observations.

IMAvatar [43] combines 3D morphable-model parameters for pose and expression control using neural blendshapes and skinning fields to produce animatable head avatars. ImFace [41] disentangles identity and expression using two deformation fields applied to a signed distance function (SDF) template. ImFace++ [42] extends this approach with a two-stage refinement framework that improves detail preservation.

NeRFs [22] map spatial coordinates and viewing directions to radiance and density and render images via volumetric integration. For head avatars, Wang et al. [35] encode sparse views into a 3D structure-aware grid of animation codes refined by an MLP. Gafni et al. [7] integrate a low-dimensional morphable face model with a neural scene representation to obtain photorealistic, controllable avatars from monocular video. Gao et al. [9] employ multilevel voxel fields with low-dimensional expression coefficients to capture elements beyond mesh blendshapes (*e.g*. hair and accessories). INSTA [45] accelerates dynamic NeRF by embedding it around a surface representation to obtain animatable avatars from short monocular video and AvatarMAV [37] decouples appearance from motion via motion-aware neural voxel grids.

3D Gaussian splatting [16] represents 3D scenes as anisotropic Gaussian primitives, and renders them via differentiable splatting. In the context of head avatars, Rig3DGS [26] reconstructed scenes in a canonical Gaussian space and learned 3DMM-guided deformations for efficient and photorealistic animation, while HeadGaS [5] extended the representation with blendable Gaussians whose attributes adapt to expression coefficients. MeGA [33] introduced a hybrid mesh–Gaussian design, combining splats with mesh geometry for high-fidelity rendering and editable head avatars. GaussianAvatars [25] bound deformable 3D Gaussians to a parametric face mesh via a binding inheritance strategy, and SurFhead [19] replaced the 3D Gaussians with 2D Gaussian surfels [14], applying Jacobian Blend Skinning and polar decomposition, achieving state-of-the-art results in dynamic head reconstruction.

2.2. Mesh Deformation and Exaggeration

Classical mesh-based approaches realize deformations using geometry processing, e.g., Poisson/Laplacian editing and related curvature-driven deformations [17, 30, 31, 40]. For faces, mesh-based deformation and caricaturization have been explored through both geometry-driven and data-driven approaches, evolving from early parametric face models to modern neural deformation networks. Early work by Blanz and Vetter [1] introduced the 3D Morphable Face Model (3DMM), representing shape and texture as linear combinations of example faces, enabling identity and expression manipulation. In the caricature domain, Brennan [2] developed an interactive system for producing line-

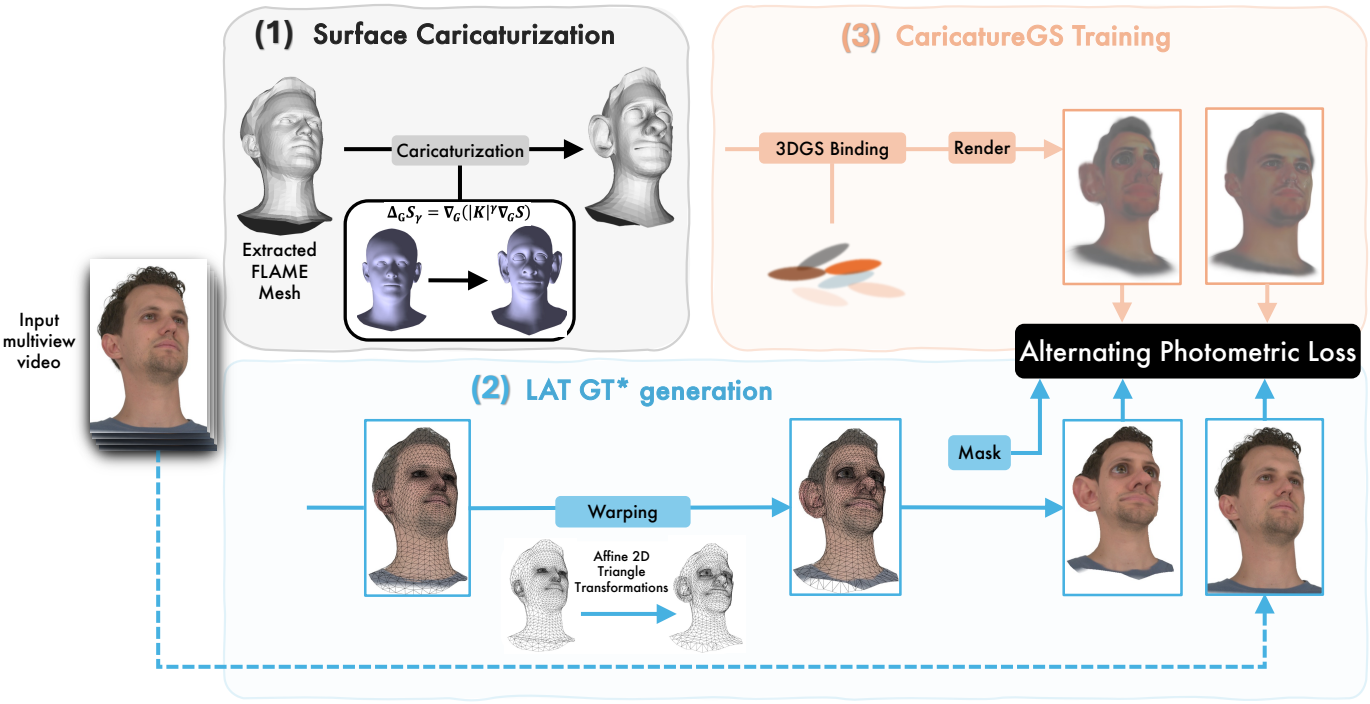


Figure 2. **CaricatureGS generation framework.** **(1)** From a subject's multi-view video, we extract a FLAME mesh and compute a curvature-driven caricature based on it. Combined with subject-specific FLAME parameters, this yields the subject's caricature mesh. **(2)** Per-triangle 2D affine transforms map the neutral mesh projection to its caricatured counterpart, warping each frame to generate pseudo–ground-truth image pairs. **(3)** Anisotropic 3D Gaussians primitives are bound to the original mesh and transformed to the caricature mesh via the corresponding 3D triangle transforms. Rendered neutral and caricature views are alternated and compared to their pseudo–ground-truth counterparts in joint optimization.

drawn caricatures by exaggerating the vector differences between the features of a subject and an average face. Eigensatz [6] used curvature maps to enhance, smooth, and transfer characteristics while preserving global structure. Later, Sela et al. [29] proposed a scale-aware Poisson-based curvature framework for surface caricaturization, exaggerating geometric features while maintaining spatial and temporal coherence.

Data-driven methods have enabled for more expressive and automated mesh exaggerations. Wu et al. [36] learned deformation patterns from artist-created examples to generate 3D caricatures from a single 2D portrait while preserving identity. Han et al. [10] introduced *DeepSketch2Face*, where a CNN infers and refines 3D face or caricature meshes from 2D sketches, while their later work *CaricatureShop* [11] combined vertex-wise Laplacian scaling with deep learning to produce photorealistic, personalized 2D caricatures from reconstructed 3D faces. Jung et al. [15] advanced this idea by using an MLP to map latent codes to 3D displacements, supporting controlled and diverse exaggerations. More recent approaches focus on style adaptation and broader correspondences. Yan et al. [38] presented an alignment-aware 3D face morphing framework with controller-based mapping for cross-species correspondence. Olivier et al. [23] explored GAN-based style transfer from scans to caricatures. Yoon et al. [39] proposed *LeGO*, a one-shot method that fine-tunes a surface deformation network to replicate a target style. An additional line of work that can be adapted to facial exaggeration is the generative line, exemplified by Diffusion- and GAN-based 3DGS editors [4, 20, 34], which operate primarily on appearance while leaving the underlying geometry unchanged.

3. Method

Here, we introduce a method for creating controllable photorealistic caricaturizations of human faces with 3DGS. Our pipeline, illustrated in Fig. 2, begins with a multiview video of a subject, from which we extract a FLAME-fitted mesh. In Sec. 3.1, we describe how we deform the geometry to obtain a caricaturized mesh. To supervise 3DGS training, we generate pseudo–ground-truth caricature images (GT*) using a 2D warping scheme (Sec. 3.2). The Gaussian primitives are then rigged to both the neutral and caricatured meshes and optimized by minimizing alternating photometric losses between their renders, the original frames, and the corresponding GT* images (Sec. 3.3). Finally, we demonstrate that this single shared Gaussian set, although trained

only on these two image domains, supports real-time rendering across a continuous range of exaggeration levels via surface interpolation and enables region-specific edits (Sec. 3.4).

3.1. Surface Caricaturization

Starting from the temporally consistent FLAME mesh obtained by fitting the landmarks [32], we apply a curvature-driven deformation that exaggerates facial geometry. Since the mesh maintains consistent vertex correspondences across frames, these deformations preserve temporal coherence. To implement this deformation, we formulate it as a weighted Poisson equation on the surface.

Let $S \in \mathbb{R}^3$ be a surface with metric G and Gaussian curvature $K(p)$ for $p \in S$. For $\gamma \in [0, \gamma_f]$, we define the *weighted Poisson equation*

$$\Delta_G S_\gamma \quad = \quad \nabla_G \cdot \big(w(\gamma) \nabla_G S\big). \tag{1}$$

We adopt the curvature-driven deformation model introduced by [28], whose weights are given by $w(\gamma) = |K|^\gamma$. This gives, for each γ, the following family of Poisson equations :

$$\Delta_G S_\gamma \quad = \quad \nabla_G \cdot \big(|K|^\gamma \nabla_G S\big). \tag{2}$$

In order to derive the deformed surface we solve the PDE by the following least-squares:

$$\min_{\tilde{x}} \|L\tilde{x} - b\|_A^2. \tag{3}$$

L is the *discrete Laplace–Beltrami operator*, defined as $L = A^{-1}W$, A is a diagonal area matrix, W is the classic *cotangent weight matrix* and $b = \nabla_G \cdot \big(|K|^\gamma \nabla_G(x)\big)$. The weighted norm is defined as $\|F\|_A^2 = \mathrm{trace}(F^T A F)$. We denote by S_γ the solution of the weighted Poisson equation in equation 2.

To accommodate open surfaces, where the Gaussian curvature may be ill defined on ∂S or to allow precise user-controlled exaggerations as discussed in Sec. 3.4, we impose boundary conditions on the selected vertices, namely:

$$\min_{\tilde{x} \in \mathbb{R}^n} \ \|L\tilde{x} - b\|_A^2 \quad \text{s.t.} \quad B\tilde{x} = x^*, \tag{4}$$

where $B \in \{0,1\}^{m \times n}$ selects the rows corresponding to the set of vertices and x^* are the prescribed boundary positions. The same constrained system is solved independently for the y and z coordinates.

An example of the resulting mesh deformation is illustrated in part **(1)** of Fig. 2.

3.2. GT* Generation via Local Affine Transforms

With these deformed surfaces, the avatar's geometry is represented in caricatured form. For photorealistic rendering, we employ mesh-rigged 3DGS, detailed in Sec. 3.3. Since using 3DGS without caricature optimization yields poor results (Sec. 4.2), training requires ground-truth supervision images. As real caricature images do not exist, we generate pseudo–ground truth (GT*): photorealistic caricature images that preserve identity while ensuring multiview consistency.

One possible way to obtain such supervision is one-shot stylization (e.g., Zhou et al. [44]), which narrows the natural–caricature gap using a single exemplar image. However, it fails to disentangle style from pose and identity, often transferring both instead of style alone (see supplementary). We therefore propose an alternative: Local Affine Transformations (LAT), illustrated in part **(2)** of Fig. 2.

LAT exploits the shared connectivity of the neutral and deformed meshes, implying a per-triangle correspondence. Consider corresponding 3D triangles $X = \{X_1, X_2, X_3\} \in \mathbb{R}^3$ and $Y = \{Y_1, Y_2, Y_3\} \in \mathbb{R}^3$. Let $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ denote the image-plane projection, with $x_i = \pi(X_i)$ and $y_i = \pi(Y_i) \in \mathbb{R}^2$. Assuming $\{x_1, x_2, x_3\}$ are non-collinear, there exists a unique affine map,

$$\Phi(\mathbf{x}) = A\mathbf{x} + \mathbf{b}, \qquad A \in \mathbb{R}^{2\times 2},\ \mathbf{b} \in \mathbb{R}^2, \tag{5}$$

such that $\Phi(x) = y$. We then used these per-triangle 2D affine transformations to map color from the original image to the 2D projection of the deformed mesh. In practice, we apply an inverse warp from each target pixel back to the original image and use bilinear interpolation to avoid empty regions.

Caricature deformation can reveal regions previously self-occluded in the neutral pose or occlude regions that were visible, leaving some pixels in GT*without valid correspondences. To address this, we generate 2D triangle-level mask for occluded regions. In addition, because hair strays fall outside the mesh limits and cannot be warped reliably, we add the hair boundary to the mask. The final output is pseudo–ground truth (GT*): high-quality caricature images that preserve identity, ensure multiview consistency, and provide effective supervision for 3DGS, together with masks indicating per-pixel validity (see appendix for further details).

3.3. CaricatureGS Training

We model the avatar's appearance photorealistically using the 3D Gaussian Splatting framework [16]. Each Gaussian g_i stores local attributes: position μ_i, scale s_i, rotation r_i, opacity σ_i, and a view-dependent color c_i. At each time frame $k \in [0, N]$, the FLAME mesh $\mathcal{M} \subset \mathbb{R}^3$ is represented by triangles $\{T_j[k]\}_{j=1}^M$, where M is the number of mesh faces. To ensure spatial–temporal coherence, each Gaussian G_i is linked [25] to a specific triangle T_j by a binding index b_i, converting its local attributes to world space.

Building on this rigged Gaussian setup, SurFhead [19] used 2D Gaussian surfels [14], which represent surfaces as oriented planar Gaussian disks, and replaced Linear Blend Skinning (LBS) with Jacobian Blend Skinning (JBS) for Gaussians deformations, namely,

$$\begin{aligned} \Sigma_i^{1/2} &= \mathbf{J_b} r_i s_i, \quad \mu_i' = \mathbf{J_b}\mu_i + T_j^x \\ \text{where } \mathbf{J_b} &= \exp\left(\sum_{i\in adj} v_i \log(U_i)\right) \cdot \sum_{i\in adj} v_i P_i, \quad (6) \end{aligned}$$

where v_i are learned weights and T_j^x is the triangle's barycentric center. U_i and P_i are the rotations and stretches from decomposing the Jacobian gradient $\mathbf{J}$ via polar decomposition. Polar decomposition separates rotation and stretch, ensuring geometrically accurate Gaussian deformations (see [19] for further details).

We show that a setup originally designed for natural facial expressions can be adapted to caricature modeling by applying the deformed caricature mesh for Gaussian deformation and using GT*for 3DGS optimization. Nevertheless, training exclusively on GT*introduces occlusion-induced artifacts and limits the model to a single expression level. To overcome these limitations, we propose a joint optimization procedure that alternates supervision randomly between real video frames and their caricatured GT*counterparts, while maintaining a single shared set of Gaussians, whose rigging ensures consistent kinematics across both supervision domains. The masks introduced in Sec. 3.2 prevent supervision of Gaussians corresponding to caricature GT*pixels that cannot be reliably warped. The joint optimization scheme allows the caricatured 3DGS to learn beyond GT*by simultaneously filling occlusion-induced holes using supervision from the original frames. As further demonstrated in Sec. 5.2, this strategy effectively captures hair details for our caricature avatar, despite hair pixels being excluded from direct GT*supervision. Moreover, as explained in Sec. 3.4, it also enables the generation of intermediate caricatures at *any* level, at inference, without additional capture.

3.4. CaricatureGS Features

The joint optimization not only complements the caricature Gaussians with information absent from GT*but present in the original frames, it also provides controllability advantages during inference.

Controlling Caricature Level. After joint training at the target exaggeration level γ_f, we empirically observe that the single-rigged Gaussian set generalizes seamlessly, rendering avatars from meshes deformed for any $\gamma \in [0, \gamma_f]$ without additional optimization. However, obtaining the deformed mesh for each γ requires solving a curvature-weighted Poisson problem, which poses a runtime bottleneck and makes interactive control of caricature levels impractical. This motivates the need for a representation that can be efficiently derived from the original mesh S_0 and the precomputed caricatured mesh S_{γ_f}. We define this representation as a vertex-wise blend:

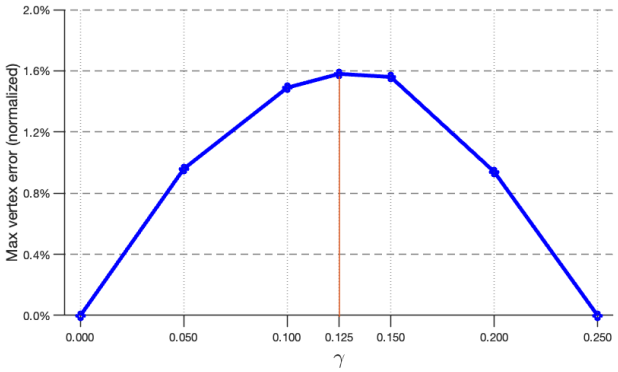


Figure 3. Parametric trend of the error with respect to γ. The error, normalized by the bounding-box diagonal of the mesh, increases from both ends of γ, reaching a negligible maximum at $\frac{\gamma_f}{2}$, where $\gamma_f = 0.25$.

$$S_{\text{blend}}(\gamma) \;=\; (1-\alpha)\, S_0 \;+\; \alpha\, S_{\gamma_f}, \quad \alpha \equiv \frac{\gamma}{\gamma_f}. \quad (7)$$

We define the residual between the approximation $S_{\text{blend}}(\gamma)$ and the exact solution $S(\gamma)$ as

$$\delta S(\gamma) \;=\; S_{\text{blend}}(\gamma) \;-\; S(\gamma). \quad (8)$$

In the supplementary material, we show that the L^2 energy of this residual can be bounded using Poincaré inequality together with the Lax-Milgram theorem given by

$$\begin{aligned} \|\delta S(\gamma)\|_{L^2} &\lesssim \tilde{C}\,\gamma(\gamma_f - \gamma)\,\|\nabla_G S_0\|_{L^2}, \\ \tilde{C} &= C_P\,(\ln|K|)^2\, e^{\max\{0,\gamma_f \ln|K|\}}, \quad (9) \end{aligned}$$

with C_P a constant.

This bound is zero at the end points $\gamma = 0, \gamma_f$, which means there is no error, as expected from (7) and maximized near $\gamma = \frac{\gamma_f}{2}$, where it remains small in practice. Empirically, we evaluate the maximal deformation error between $S_{\text{blend}}(\gamma)$ and S_γ on varying γ and different subjects, normalized by the mesh bounding-box diagonal. As shown in Fig. 3, the worst-case deviation is negligible, supporting the fidelity of the interpolation and confirming that it lies near the theoretical midpoint of the exaggeration, as predicted. This implies that, with this approximation, no additional Poisson equations need to be solved when inferring new γ values, thereby enabling full interactive control of caricature levels. In Fig. 5, we illustrate that this interpolation scheme enables a single set of Gaussians to smoothly represent shape deformations across the full range of γ.

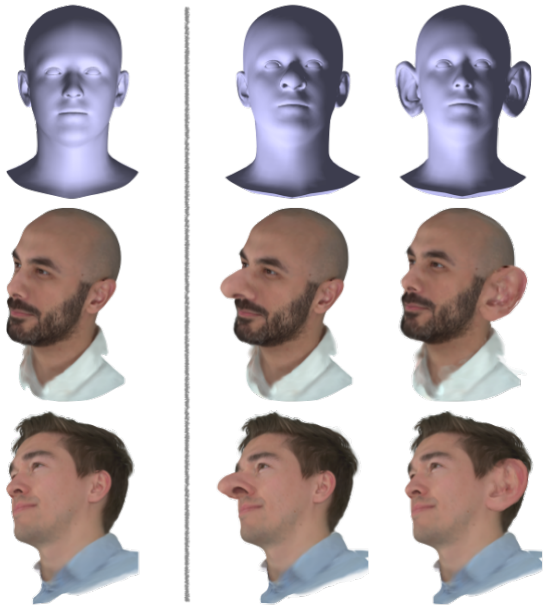


Figure 4. Visualizations of localized, semantically controlled facial exaggerations.

Localized Caricature Control. Our curvature-weighted model uses the local curvature K to generate a globally consistent caricature by solving the unconstrained Poisson equation. To target specific regions, we solve the constrained least-squares system in Eq. (4), whereby only the chosen region of interest undergoes curvature deformations, producing a smooth and localized exaggerations that blend harmonically with the rest of the face. Coupled with the training scheme in Sec. 3.3, the 3DGS, rigged to the mesh, faithfully tracks these deformations, so the same Gaussian set realizes semantically controlled exaggerations while preserving identity and global shape (see Fig. 4).

4. Experiments

We evaluate our caricaturized avatars along two main axes: (i) photorealistic rendering, (ii) identity preservation. All experiments are conducted on the NeRSemble dataset [18] and compared against the recent state-of-the-art 4D avatar reconstruction method of SurFhead [19]. Unless noted otherwise, we apply an unconstrained exaggeration with $\gamma_f = 0.25$.

4.1. Dataset

The NeRSemble dataset [18] provides a multi-view facial performance dataset captured by 16 spatially arranged, synchronized high-resolution cameras. It comprises 10 scripted sequences, 4 emotion-driven (EMO) and 6 expression-driven (EXP), plus an additional free self-reenactment sequence. For fair comparison, we adopt the same train/validation/test partition as in [19] with $120,000$ training iterations. Further implementation details are provided in the supplementary.

Method	CLIP-I ↑	CLIP-D ↑	CLIP-C ↑	DINO ↑	SD ↑
SurFhead	0.67	0.0006	0.944	0.757	0.460
Ours	0.73	0.014	0.945	0.888	0.539

Table 1. Quantitative comparison for a caricature avatar. Higher is better for all reported metrics.

4.2. Baseline

To the best of our knowledge, there are no explicit methods that construct a dynamic 3D photorealistic model from an input multi-view video. To this end, we compare with SurFhead [19] using the authors' official implementation. SurFhead achieves state-of-the-art performance in head reconstruction and reenactment and, in principle, can handle mesh deformations through JBS, making it the most suitable baseline for comparison. We train the SurFhead on the original input sequence and, at inference, we exaggerate the underlying mesh using γ_f, as elaborated in Sec. 2.2, thereby driving the Gaussians to represent a caricaturized avatar.

4.3. Metrics

Quantitative evaluation of caricature models is inherently challenging due to their under-constrained nature and the lack of ground-truth images. We use the following metrics for evaluation:

- **CLIP-I** (Image–Prompt Similarity) [13]: Cosine similarity between the rendered image and text in CLIP space.
- **CLIP-D** (Directional Similarity) [8]: Measures the change between source and edited images against the change between source and edited prompts.
- **CLIP-C** (Spatial Consistency): Following [12], we report CLIP image alignment between adjacent novel views of image embeddings along a novel trajectory.
- **DINO** (Identity/Structure Consistency): Following [44], we extract DINO [3] features from the renders and the corresponding original test frames and compute the cosine similarity of the embeddings.
- **SD** (Score Distillation): Inspired by DreamFusion [24], we define the reference-free metric as,

$$\mathrm{SD} = 1 - \frac{1}{BTN} \sum_{b,t,n}^{B,T,N} \frac{\left\| \epsilon_\theta\left(x_t^{(b,t,n)},\, t\right) - \epsilon_{b,t,n} \right\|_2^2}{\|\epsilon_{b,t,n}\|_2^2}. \quad (10)$$

  where $\epsilon_\theta(x_t, t)$ is the noise predicted by the diffusion model [27] at time step t, ϵ is the true noise, and B, T, N refer to the image count, time step, and seed number, respectively. Higher SD indicates that the rendered image is more consistent with the training distribution of the diffusion model, which is intended to approximate the natural image distribution.

Text prompts are provided in the appendix. Together, these metrics evaluate: (i) how well the renders reflect the carica-

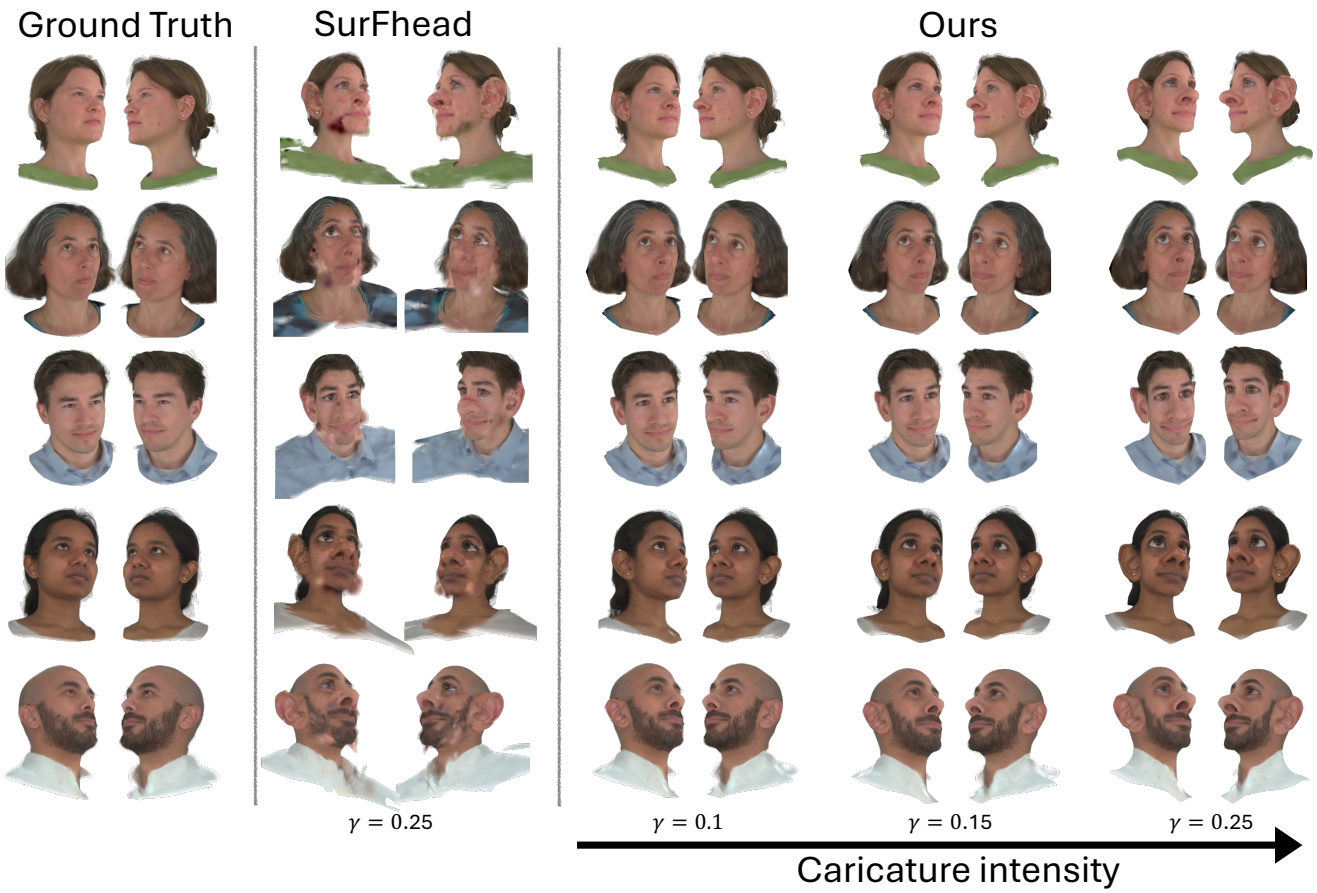


Figure 5. Rendering results from our pipeline [19]. **SURFHEAD:** Caricature generation by first reconstructing an avatar with the state-of-the-art SURFHEAD model [19], followed by mesh exaggeration. **Ours:** Renderings across different caricature intensities. Our approximation-based control interpolates smoothly along the caricature intensity axis while preserving visual fidelity.

ture intent (CLIP-I, CLIP-D, SD), (ii) identity preservation and the extent to which exaggerations remain localized to caricaturization (DINO, CLIP-D), and (iii) consistency of generated views across novel trajectories (CLIP-C).

4.4. Results

Fig. 5 presents side-by-side renderings at the target exaggeration level γ_f for our method and the baseline. Our approach maintains subject identity while delivering natural, visually pleasing exaggerations that remain consistent across views, and reduces the distortions visible in the baseline. The figure further illustrates caricature-level controllability by varying γ from 0 to γ_f, demonstrating continuous control and showing that the approximation in Sec. 3.4 successfully supports intermediate exaggeration levels.

For quantitative evaluation, we conduct a comprehensive comparison using the metrics in Sec. 4.3. As summarized in Tab. 1, our method consistently surpasses the baseline across all measures, demonstrating that the learned edits faithfully capture the intended caricature while preserving both identity and view-consistency.

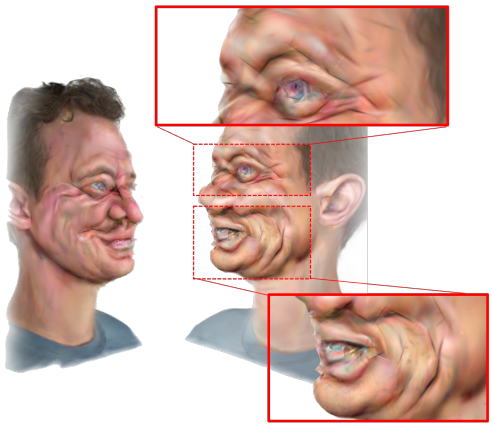
4.5. Diffusion Based Editing

As an additional baseline, we adapt a diffusion-driven, text-guided, mesh-free 3DGS editor [4] for caricaturization. Using the authors' implementation, we run 5,000 optimization steps per prompt on multiview images of a subject, guided by ControlNet-Pix2Pix. Fig. 6a presents a global edit, while Fig. 6b shows a local edit, manually masked for face and nose, respectively. While the edits appear visually plausible in individual views, it is evident that, unlike our method, this baseline suffers from (i) geometry drift, (ii) unstable, view-dependent specularities, and (iii) poor multi-view coherence.

5. Ablations

5.1. Alternated Training

In this subsection, we demonstrate that training with GT*, generated using LAT, is essential for controlling the caricaturization level. As discussed in Sec. 4.4, training only on input images fails to generalize: rendering with a caricatured mesh yields heavily degraded outputs. In the supple-



(a) Edit instruction: "Turn him into a realistic caricature." The result exhibits skin-tone shifts and specular degradation.

(b) Edit instruction: "Make his nose bigger." The geometry falls apart and color inconsistencies appear across views.

Figure 6. GaussianEditor [34] caricaturization attempts. **(a)** Global edit. **(b)** Local semantic edit. Both reveal degraded geometry and appearance fidelity, particularly in novel views.

mentary, we show that training solely with GT* also fails: neutral renders appear unrealistic, with distorted Gaussian structures. These complementary failures underscore the necessity of alternating both forms of supervision for effective caricaturization control.

5.2. Mask

Due to the nature of GT* generation, certain fine details, most notably hair, are often misrepresented during the caricature stage. To address this, we identify hair regions of the mesh and freeze the corresponding Gaussian parameters with a suitable mask during GT* supervision iterations, thereby preventing updates in those regions when the caricature is rendered (see Sec. 3.2). Fig. 7 illustrates the effect: on the left, hair regions are masked and remain frozen, whereas on the right they are unfrozen and allowed to train freely, resulting in unnaturally plastic-looking hair.

6. Limitations

While our method provides a powerful framework for photorealistic 3D caricaturization, several limitations remain.

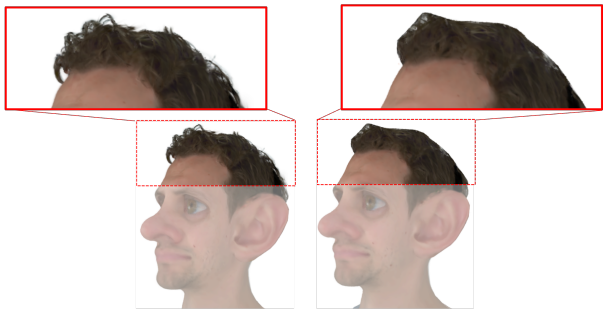


Figure 7. Ablation on hair masking. Without masking, GT* introduces visible artifacts in hair regions. Masking and freezing Gaussians associated with hair during GT* supervision effectively prevents these artifacts.

Although our approach improves upon the baseline, residual specularity artifacts persist, and small eyelid inaccuracies—amplified by over-stretching in LAT, become visually noticeable. This effect also extends to hair: training Caricature 3DGS hair with input-view supervision alone (without GT*) substantially alleviates the issue. However, in some cases, we observe slight over-smoothing of the hair. Qualitative examples of these effects are provided in the supplementary material. Finally, the deformed FLAME mesh does not fully span the space of facial expressions. For instance, eyelid closure in caricatured results is imperfect: eyes that should be completely shut under certain expressions often remain slightly open, leading to misrepresentations of eyelid geometry in the final caricature.

7. Discussion

This work demonstrates that curvature-driven geometric deformation and mesh-rigged 3D Gaussian Splatting (3DGS) can be combined into a single, controllable avatar model that remains photorealistic under large exaggerations. The key is a training scheme that alternates supervision between real views and generated pseudo–ground-truth caricature views, produced using per-triangle Local Affine Transformations (LAT) with reliability masks. One Gaussian set is capable of jointly learning both natural and caricatured appearance while retaining identity and expression. Prior work indicates that deliberate shape exaggeration can amplify discriminative geometric cues for recognition [28]. Looking ahead, we hypothesize that integrating our controllable exaggeration as a plug-in augmentation within face-recognition pipelines could improve robustness to pose and expression variability. Finally, coupling our geometry-grounded deformations with diffusion-based editors may enable semantically guided edits that are both photorealistic and extend beyond appearance-only changes to joint control of shape and appearance.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[2] Susan E. Brennan. Caricature Generator: The Dynamic Exaggeration of Faces by Computer. *Leonardo*, 18(3):170–178, 1985. Publisher: The MIT Press.

[3] Mathieu Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[4] Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. GaussianEditor: Swift and Controllable 3D Editing with Gaussian Splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21476–21485, Seattle, WA, USA, 2024. IEEE.

[5] Helisa Dhamo, Yinyu Nie, Arthur Moreau, Jifei Song, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. HeadGaS: Real-Time Animatable Head Avatars via 3D Gaussian Splatting. *arXiv e-prints*, art. arXiv:2312.02902, 2023.

[6] Michael Eigensatz, Robert W. Sumner, and Mark Pauly. Curvature-Domain Shape Processing. *Computer Graphics Forum*, 27(2):241–250, 2008. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2008.01121.x.

[7] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. *arXiv e-prints*, art. arXiv:2012.03065, 2020.

[8] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.

[9] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing Personalized Semantic Facial NeRF Models From Monocular Video. *arXiv e-prints*, art. arXiv:2210.06108, 2022.

[10] Xiaoguang Han, Chang Gao, and Yizhou Yu. DeepSketch2Face: a deep learning based sketching system for 3D face and caricature modeling. *ACM Transactions on Graphics*, 36(4):1–12, 2017.

[11] Xiaoguang Han, Kangcheng Hou, Dong Du, Yuda Qiu, Yizhou Yu, Kun Zhou, and Shuguang Cui. CaricatureShop: Personalized and Photorealistic Caricature Sketching, 2018. arXiv:1807.09064 [cs].

[12] Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. *arXiv e-prints*, art. arXiv:2303.12789, 2023.

[13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *arXiv e-prints*, art. arXiv:2104.08718, 2021.

[14] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24*, pages 1–11, 2024. arXiv:2403.17888 [cs].

[15] Yucheol Jung, Wonjong Jang, Soongjin Kim, Jiaolong Yang, Xin Tong, and Seungyong Lee. Deep Deformable 3D Caricatures with Learned Shape Control. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, pages 1–9, 2022. arXiv:2207.14593 [cs].

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

[17] ByungMoon Kim and Jarek Rossignac. Geofilter: Geometric selection of mesh filter parameters. *Comput. Graph. Forum*, 24:295–302, 2005.

[18] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. NeRSemble: Multi-view Radiance Field Reconstruction of Human Heads. *ACM Transactions on Graphics*, 42(4):1–14, 2023. arXiv:2305.03027 [cs].

[19] Jaeseong Lee, Taewoong Kang, Marcel C. Bühler, Min-Jung Kim, Sungwon Hwang, Junha Hyung, Hyojin Jang, and Jaegul Choo. SurFhead: Affine Rig Blending for Geometrically Accurate 2D Gaussian Surfel Head Avatars, 2024. arXiv:2410.11682 version: 1.

[20] Guohao Li, Hongyu Yang, Yifang Men, Di Huang, Weixin Li, Ruijie Yang, and Yunhong Wang. Generating Editable Head Avatars with 3D Gaussian GANs, 2024. arXiv:2412.19149 [cs].

[21] 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.

[22] 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. *CoRR*, abs/2003.08934, 2020.

[23] Nicolas Olivier, Glenn Kerbiriou, Ferran Argelaguet Sanz, Quentin Avril, Fabien Danieau, Philippe Guillotel, Ludovic Hoyet, and Franck Multon. Study on Automatic 3D Facial Caricaturization: From Rules to Deep Learning. *Frontiers in Virtual Reality*, 2:1–15, 2022.

[24] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D Diffusion, 2022. arXiv:2209.14988 [cs].

[25] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians, 2024. arXiv:2312.02069 [cs].

[26] Alfredo Rivero, ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. Rig3DGS: Creating Controllable Portraits from Casual Monocular Videos. *arXiv e-prints*, art. arXiv:2402.03723, 2024.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[28] Matan Sela, Yonathan Aflalo, and Ron Kimmel. Computational caricaturization of surfaces. *Computer Vision and Image Understanding*, 141:1–17, 2015.

[29] Matan Sela, Yonathan Aflalo, and Ron Kimmel. Computational caricaturization of surfaces. *Computer Vision and Image Understanding*, 141:1–17, 2015.

[30] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Proceedings of EUROGRAPHICS/ACM SIGGRAPH Symposium on Geometry Processing*, pages 109–116, 2007.

[31] Olga Sorkine, Daniel Cohen-Or, Yaron Lipman, Marc Alexa, Christian Rössl, and Hans-Peter Seidel. Laplacian surface editing. In *Proceedings of the EUROGRAPHICS/ACM SIGGRAPH Symposium on Geometry Processing*, pages 179–188. ACM Press, 2004.

[32] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. *arXiv e-prints*, art. arXiv:2007.14808, 2020.

[33] Cong Wang, Di Kang, He-Yi Sun, Shen-Han Qian, Zi-Xuan Wang, Linchao Bao, and Song-Hai Zhang. MeGA: Hybrid Mesh-Gaussian Head Avatar for High-Fidelity Rendering and Head Editing. *arXiv e-prints*, art. arXiv:2404.19026, 2024.

[34] Junjie Wang, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, and Qi Tian. GaussianEditor: Editing 3D Gaussians Delicately with Text Instructions, 2024. arXiv:2311.16037 [cs].

[35] Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhöfer. Learning Compositional Radiance Fields of Dynamic Human Heads. *arXiv e-prints*, art. arXiv:2012.09955, 2020.

[36] Qianyi Wu, Juyong Zhang, Yu-Kun Lai, Jianmin Zheng, and Jianfei Cai. Alive Caricature from 2D to 3D, 2018. arXiv:1803.06802 [cs].

[37] Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. AvatarMAV: Fast 3D Head Avatar Reconstruction Using Motion-Aware Neural Voxels. *arXiv e-prints*, art. arXiv:2211.13206, 2022.

[38] Xirui Yan, Zhenbo Yu, Bingbing Ni, and Hang Wang. Cross-Species 3D Face Morphing via Alignment-Aware Controller. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):3018–3026, 2022.

[39] Soyeon Yoon, Kwan Yun, Kwanggyoon Seo, Sihun Cha, Jung Eun Yoo, and Junyong Noh. LeGO: Leveraging a Surface Deformation Network for Animatable Stylized Face Generation with One Example, 2024. arXiv:2403.15227 [cs] version: 1.

[40] Yizhou Yu, Kun Zhou, Dong Xu, Xiaohan Shi, Hujun Bao, Baining Guo, and Heung-Yeung Shum. Mesh editing with poisson-based gradient field manipulation. *ACM Trans. Graph.*, 23(3):644–651, 2004.

[41] Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. ImFace: A Nonlinear 3D Morphable Face Model with Implicit Neural Representations. *arXiv e-prints*, art. arXiv:2203.14510, 2022.

[42] Mingwu Zheng, Haiyu Zhang, Hongyu Yang, Liming Chen, and Di Huang. ImFace++: A Sophisticated Nonlinear 3D Morphable Face Model with Implicit Neural Representations. *arXiv e-prints*, art. arXiv:2312.04028, 2023.

[43] 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. *arXiv e-prints*, art. arXiv:2112.07471, 2021.

[44] Yang Zhou, Zichong Chen, and Hui Huang. Deformable One-shot Face Stylization via DINO Semantic Guidance, 2024. arXiv:2403.00459 [cs].

[45] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant Volumetric Head Avatars. *arXiv e-prints*, art. arXiv:2211.12499, 2022.