STRAINAVATAR: STRAIN-DRIVEN HEAD AVATARS WITH EXPLICIT FACIAL SKIN DEFORMATION

Anonymous authors
Paper under double-blind review



Figure 1: Head avatars reconstructed by StrainAvatar from multi-view RGB. Each panel pairs a rendering with its normals. The bottom row zooms into a wrinkle region of the same pair.

ABSTRACT

Head avatars reconstructed from multi-view video achieve photorealistic rendering, yet expression-dependent wrinkles are typically synthesized in appearance or modeled as directly predicted displacements, leaving their geometric cause unmodeled. We introduce StrainAvatar, a head avatar that reconstructs wrinkle-scale facial deformation as explicit geometry from multi-view RGB alone. Our key idea is to factor wrinkle formation into a shared mechanical cause and an identity-specific geometric response. Facial compression, analytically characterized by the strain of the tracked deformation, provides an identity-agnostic signal that generalizes across expressions, while a subject-specific network learns only how the skin responds to this signal. We realize this factorization by predicting per-triangle Jacobian correctives from the strain field and integrating them into a coherent corrected mesh through a Poisson solve. Because a zero corrective exactly recovers the tracked mesh, large-scale facial motion remains deterministic and learning is restricted to wrinkle-scale deformation. We further transport Gaussian primitives using the deformation gradient of the corrected mesh, making appearance explicitly inherit the recovered surface deformation. Experiments show rendering quality comparable to recent head avatars, improved surface-normal accuracy, and consistent reductions in mesh error against independent multi-view stereo depth, qualitatively recovering expression-dependent wrinkle structure as explicit, animatable geometry.

1 INTRODUCTION

Creating photorealistic and animatable head avatars from multi-view video has been a longstanding problem in computer vision and graphics, enabling numerous applications in immersive telepres-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ence Lombardi et al. (2018), digital humans, and visual effects. These applications demand avatars that not only render realistically from arbitrary viewpoints but also reproduce facial deformation faithfully under novel expressions. Human observers are acutely sensitive to facial appearance Saito et al. (2024), and recent advances in neural rendering Lombardi et al. (2018); Thies et al. (2019); Kirschstein et al. (2023) have pushed avatar appearance toward photographic realism. Yet this progress has largely been achieved by improving how facial detail is rendered rather than how it is geometrically formed. In particular, expression-dependent wrinkles, the fine folds produced by facial deformation, are rarely recovered as explicit geometry.

Recent head avatars combine a tracked parametric face model Li et al. (2017) with 3D Gaussian splatting Kerbl et al. (2023). The parametric model supplies the geometry and the animation control, and the primitives are rigged to its mesh so that the avatar follows the tracked expression Qian et al. (2024); Shao et al. (2024); Xiang et al. (2024); Lee et al. (2026a). Throughout these developments, however, the underlying geometry has largely remained the tracked mesh itself. Expression-dependent detail is instead synthesized through appearance: the dominant strategy conditions primitive attributes on an expression code Wang et al. (2025c); Li et al. (2025); Aneja et al. (2025), leaving fine geometry only implicitly constrained by photometric supervision. Even methods that explicitly deform the geometry remain constrained by the smooth morphable-model representation Lee et al. (2025). A complementary direction refines facial geometry with displacement maps Feng et al. (2021); Wang et al. (2025a); Lee et al. (2026b), but recovering wrinkle-scale geometry from images remains difficult: locally predicted displacement is weakly constrained by photometric supervision, while regularization needed for stable optimization tends to suppress precisely the high-frequency folds of interest. Consequently, high-fidelity wrinkle geometry, for faces and garments alike, has typically relied on dedicated 3D capture Ma et al. (2008); Bickel et al. (2008); Guo et al. (2024); Zhang et al. (2023); Wang et al. (2025b). Recovering such expression-dependent deformation as explicit, animatable geometry from multi-view RGB alone therefore remains an open problem.

Our key departure is to predict neither wrinkles nor displacements directly. Instead, we factor wrinkle formation into a mechanically determined cause and an identity-specific geometric response. Facial skin wrinkles under compression Bickel et al. (2008), as when raising the eyebrows or frowning. This compression can be measured analytically from the strain of the tracked facial deformation. Because strain describes how the surface is being deformed, rather than which expression generated it, it provides a local, identity-agnostic signal that is available for any tracked expression. The resulting wrinkles are identity-specific in their location, orientation, amplitude and wavelength. We therefore use the analytic strain field to represent the shared mechanical cause and learn only the subject-specific response. This differs from predicting fine geometry from an expression or pose code: the network is conditioned on the local deformation state that drives wrinkle formation, while the coarse facial motion itself remains analytically determined.

We introduce **StrainAvatar**, a strain-driven head avatar with explicit facial skin deformation that reconstructs wrinkle-scale expression geometry from multi-view RGB (Fig. 1). Given a tracked expression, we analytically compute its per-triangle deformation gradient and corresponding skin strain. An identity-specific network then predicts *Jacobian correctives from this strain field rather than predicting the facial deformation itself*. The correctives are composed with the tracked deformation gradient and integrated through a Poisson solve into a coherent corrected mesh, following the representation of Neural Jacobian Fields Aigerman et al. (2022). Gaussian primitives are then bound to the corrected mesh and transported by its deformation gradient, so the rendered appearance inherits the recovered wrinkle geometry.

- We introduce a *strain-driven factorization of facial deformation* that separates an analytically determined deformation signal from a learned identity-specific wrinkle response, enabling wrinkle-scale expression geometry to be recovered explicitly from multi-view RGB images.

- We formulate the learned response as *per-triangle Jacobian correctives* composed with the tracked deformation and integrated through a canonical Poisson solve, such that zero correction exactly recovers the tracked mesh and learning is restricted to deformation missing from the tracker.

- We present a *deformation-consistent Gaussian appearance* representation that propagates the recovered surface deformation to Gaussian positions and covariances while maintaining surface alignment, directly coupling rendered appearance to the explicit wrinkle geometry.

2 Related Work

Photorealistic Head Avatars. Head avatars have been modeled with tracked meshes and view-dependent textures Lombardi et al. (2018) or with deformable point clouds Zheng et al. (2023). Recent methods build on 3D Gaussian splatting Kerbl et al. (2023) with the primitives bound to a tracked parametric face mesh Li et al. (2017). GaussianAvatars Qian et al. (2024), SplattingAvatar Shao et al. (2024) and FlashAvatar Xiang et al. (2024) attach the primitives to the surface of the tracked mesh with triangle frames, barycentric coordinates and learned offsets, NPGA Giebenhain et al. (2024) conditions the dynamics on the expression space of a neural parametric head model, SurFhead Lee et al. (2025) rigs 2D surfels with the Jacobians of the tracked mesh, and TexAvatars Lee et al. (2026a) places the primitives in the UV space of the morphable model and predicts their attributes with a decoder. Expression-dependent detail in these avatars is carried by appearance. HeadGaS Dhamo et al. (2024) and ScaffoldAvatar Aneja et al. (2025) decode expression-conditioned features into per-primitive color and opacity, Wang et al. (2025c) decode dynamic texture features into opacity offsets, and TeGA Li et al. (2025) and Yang et al. (2026) darken the primitives to increase the visibility of the wrinkles. TexAvatars additionally conditions its decoders on a per-frame embedding that the image animation model EMOPortraits Drobyshev et al. (2024) extracts from the images. A separate group refines the surface itself with a displacement map. DECA Feng et al. (2021) regresses expression-dependent detail into a UV displacement map applied at render time, MeGA Wang et al. (2025a) refines the facial mesh with a displacement map supervised by multi-view stereo depth and keeps wrinkle detail in a dynamic texture, and URHead Lee et al. (2026b) optimizes displacement and appearance maps in a shared UV parameterization with pseudo ground truth normals. Monocular avatars place the detail in appearance as well, with a dynamic texture over static facial geometry Grassal et al. (2022) or a smooth learned blendshape model Zheng et al. (2022); Bharadwaj et al. (2023). In these mesh-rigged approaches, the underlying surface generally remains the tracked morphable-model mesh, whose low-dimensional linear representation limits fine expression-dependent geometry Lee et al. (2025). Consequently, expression details visible in the rendered images are not explicitly represented in the geometry.

Mesh Deformation Models. A second line of work represents deformation by per-triangle Jacobians in place of vertex offsets. Deformation transfer Sumner & Popović (2004) computes the deformation gradient of every triangle of a source mesh and applies it to a target mesh by solving a linear system for the vertex positions. Neural Jacobian Fields Aigerman et al. (2022) predict the Jacobian of every triangle with an MLP conditioned on a global code and recover the deformed mesh with a Poisson solve, NFR Qin et al. (2023) builds a deformation autoencoder on per-triangle Jacobians and learns facial rigging from registered scans, and SurFhead Lee et al. (2025) applies the Jacobians of the tracked mesh to the primitives of a head avatar and blends the Jacobians of adjacent triangles. JacobianAvatar Won et al. (2026) reconstructs full-body avatars from monocular video with pose-conditioned Jacobian fields guided by geometric pseudo ground truth from a foundation model. The strain of the surface has been used as the signal that drives fine deformation. Bickel et al. (2008) learn the relation between sparse measurements of skin strain and wrinkle formation in pose space, and Guo et al. (2024) condition a diffusion model on the Green–Lagrange strain field of a coarse garment mesh and generate wrinkle deformation from registered clothing scans. Wrinkle geometry at this scale has otherwise been acquired by structured-light scanning and driven by motion capture markers Ma et al. (2008) or learned as pose-dependent displacements from 3D scans of clothed subjects Zhang et al. (2023); Wang et al. (2025b). The deformation in these methods is predicted from a global code or a pose vector and supervised by registered scans or dedicated capture, so it is not recovered from multi-view images alone.

3 Method

As shown in Fig. 2, we reconstruct an animatable head avatar with expression-dependent wrinkles represented as explicit mesh deformation through two stages coupled by the corrected geometry. The geometry stage recovers wrinkle-scale deformation missing from the tracked mesh. We analytically derive a skin-strain signal from the tracked deformation and use it to drive an identity-specific network that predicts corrective Jacobians, which are integrated into explicit wrinkle geometry (Sec. 3.2). The appearance stage then constructs a photorealistic Gaussian representation whose primitives are transported by the corrected surface deformation, explicitly coupling appearance to the recovered wrinkle geometry (Sec. 3.3). Both stages are trained from multi-view RGB without explicit 3D geometry supervision (Sec. 3.4). We track a FLAME Li et al. (2017) mesh for every frame and denote its vertices by V_c and the facial triangles by f, over which wrinkle deformation is modeled. All deformations are defined relative to the canonical FLAME mesh.

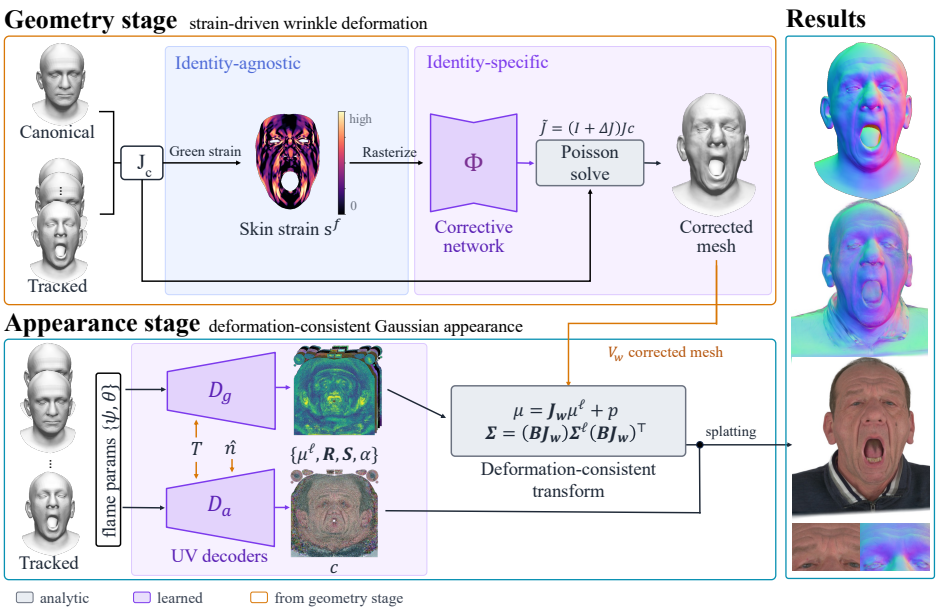


Figure 2: **Method overview.** The geometry stage derives the skin strain of every triangle from the canonical and the tracked mesh and rasterizes it over the UV map. The corrective network predicts a Jacobian corrective from this strain and a Poisson solve integrates the corrected Jacobians into the corrected mesh. The appearance stage predicts the primitive attributes in UV space and transports every primitive by the deformation gradient of the corrected mesh before splatting.

3.1 Preliminaries

Neural Jacobian Fields (NJF). Aigerman et al. (2022) represent a mesh deformation by assigning a Jacobian $\boldsymbol{J}^f \in \mathbb{R}^{3\times3}$ to each triangle f and recovering the deformed mesh vertices $\boldsymbol{V}$ as the least-squares solution

$$\boldsymbol{V}^* = \arg\min_{\boldsymbol{V}} \sum_f |f| \left\| \nabla_f \boldsymbol{V} - \boldsymbol{J}^f \right\|_F^2, \tag{1}$$

where ∇_f is the per-triangle gradient operator and $|f|$ is the triangle area. Since the assigned Jacobians need not form an integrable field, the solve projects them onto a single coherent surface. This reduces to a Poisson system whose Laplacian depends only on the mesh and can therefore be prefactorized. In NJF, the Jacobians are predicted by an MLP conditioned on a global target-shape code. We adopt this Jacobian-based representation, but unlike NJF, which predicts the target deformation itself, we predict only a strain-driven corrective on top of the analytically known tracked deformation (Sec. 3.2).

3.2 Strain-Driven Wrinkle Deformation

We factor the deformation of each triangle into an analytically determined tracked deformation and a learned corrective that captures the identity-specific wrinkle response.

Skin strain signal. Let $\boldsymbol{J}_c^f \in \mathbb{R}^{3\times3}$ denote the deformation gradient of triangle f from the canonical mesh to the tracked mesh $\boldsymbol{V}_c$. We derive $\boldsymbol{J}_c^f$ analytically from the triangle edge vectors following deformation transfer Sumner & Popović (2004). Restricting it to an orthonormal tangent basis $\boldsymbol{W}_f \in \mathbb{R}^{3\times2}$ of the canonical triangle gives $\boldsymbol{D}^f = \boldsymbol{J}_c^f \boldsymbol{W}_f \in \mathbb{R}^{3\times2}$, from which the Green–Lagrange strain tensor Sifakis & Barbic (2012) follows as

$$\boldsymbol{E}^f = \tfrac{1}{2}\left((\boldsymbol{D}^f)^\top \boldsymbol{D}^f - \boldsymbol{I}\right) \in \mathbb{R}^{2\times2}. \tag{2}$$

Because $\boldsymbol{E}^f$ is invariant to rigid motion, it isolates the local in-plane deformation of the triangle. To obtain a spatially consistent signal over the face, we express the strain tensor in the UV-aligned basis with the rotation $\boldsymbol{Q}_f \in \mathbb{R}^{2\times2}$ that maps the tangent basis $\boldsymbol{W}_f$ onto the UV axes of triangle f, and stack its three independent entries into the strain signal,

$$s^f = \left(E^f_{uu}, E^f_{vv}, E^f_{uv}\right), \qquad E^f_{ab} = \left(\boldsymbol{Q}_f^\top \boldsymbol{E}^f \boldsymbol{Q}_f\right)_{ab}, \tag{3}$$

where E^f_{uu} and E^f_{vv} are the signed normal strains along the two UV axes and E^f_{uv} is the in-plane shear. This signal characterizes the resulting local deformation state rather than directly encoding

the expression parameters, head pose, or camera viewpoint. We rasterize s^f over the face region of the FLAME UV map to obtain a strain map $\boldsymbol{T} \in \mathbb{R}^{3\times R\times R}$, where R denotes the UV map resolution. Appendix B shows the strain field.

Jacobian correctives. We predict a corrective Jacobian field from the strain map $\boldsymbol{T}$. Because the corrective network Φ is convolutional over the UV map, each corrective depends on the strain of its neighborhood. The network is trained per identity and predicts a bounded corrective field in the local tangent frame of each triangle, so its output is invariant to the global head orientation. For each triangle f, the corrective is bilinearly interpolated at its UV coordinates and rotated to world coordinates as

$$\Delta \boldsymbol{J}^f = \boldsymbol{R}_f \, \Phi(\boldsymbol{T})[f] \, \boldsymbol{R}_f^\top, \tag{4}$$

where $\boldsymbol{R}_f \in \mathbb{R}^{3\times 3}$ is the rotation from the local tangent frame of triangle f to world coordinates and $\Phi(\boldsymbol{T})[f] \in \mathbb{R}^{3\times 3}$ denotes the interpolated corrective. The strain map is the only input to Φ, while identity-specific spatial variation is encoded by learned bias maps at each encoder level. The architecture of Φ is given in Appendix C.

We compose the resulting corrective with the analytically determined tracked deformation gradient,

$$\tilde{\boldsymbol{J}}^f = \left(\boldsymbol{I} + \Delta \boldsymbol{J}^f\right) \boldsymbol{J}_c^f. \tag{5}$$

$\boldsymbol{J}_c^f$ maps the canonical triangle to its tracked configuration and $\boldsymbol{I} + \Delta \boldsymbol{J}^f$ adds the wrinkle-scale response in the tracked one. This composition is analogous to corrective blendshapes in traditional facial rigging Lewis et al. (2014), but operates in the Jacobian domain and is driven by skin strain.

Canonical Poisson integration. We recover the corrected mesh $\boldsymbol{V}_w$ from the corrected Jacobian field $\tilde{\boldsymbol{J}}^f$ of Eq. 5 by the Poisson solve of Eq. 1. We denote the Jacobians of $\boldsymbol{V}_w$ by $\boldsymbol{J}_w$. We restrict the solve to the face region and anchor its boundary to the tracked mesh,

$$\boldsymbol{V}_w = \arg\min_{\boldsymbol{V}} \sum_f |f| \left\| \nabla_f \boldsymbol{V} - \tilde{\boldsymbol{J}}^f \right\|_F^2 \quad \text{s.t.} \quad \boldsymbol{V}\big|_\partial = \boldsymbol{V}_c\big|_\partial, \tag{6}$$

where ∂ denotes the boundary vertices of the face region. This constraint keeps the corrected face continuously attached to the surrounding tracked head and is detailed in Appendix A.

Unlike NJF, which constructs the gradient operator for each source mesh, we define the operator on the canonical mesh and prefactorize it once for training and animation. Since $\boldsymbol{J}_c^f$ is the gradient of the tracked mesh under this operator, the solve recovers $\boldsymbol{V}_w = \boldsymbol{V}_c$ exactly when $\Delta \boldsymbol{J}^f = \boldsymbol{0}$ for all f and every deviation from the tracked mesh arises from the learned correctives.

3.3 Deformation-Consistent Gaussian Appearance

Our appearance stage represents the avatar with 3D Gaussian primitives Kerbl et al. (2023) and adopts the Gaussian attribute maps in the FLAME UV space of Li et al. (2025); Lee et al. (2026a). The position and the covariance of each primitive are attached to its triangle, expressed in the canonical configuration and transformed to world space by the deformation of that triangle. We transport each primitive with the corrected mesh so that the rendered appearance inherits the recovered wrinkle deformation. The position μ and the covariance $\boldsymbol{\Sigma}$ are transformed by the deformation gradient $\boldsymbol{J}_w$ and the covariance is re-aligned with the deformed surface normal.

Attribute decoders. A geometry decoder $\mathcal{D}_g$ and an appearance decoder $\mathcal{D}_a$ predict the Gaussian attributes from the FLAME expression parameters ψ and pose parameters θ. We additionally condition both decoders on the strain map $\boldsymbol{T}$ of Eq. 3, the input of the corrective network Φ. While ψ and θ provide global expression and pose conditioning, $\boldsymbol{T}$ provides the local skin strain that drives wrinkle formation. Appendix F details the decoders. The decoders produce the Gaussian attribute maps in the FLAME UV space,

$$\boldsymbol{G} = \mathcal{D}_g(\psi, \theta, \boldsymbol{T}), \quad \{\mu^\ell, \boldsymbol{R}, \boldsymbol{S}, \alpha\} \subset \boldsymbol{G}, \qquad \boldsymbol{A} = \mathcal{D}_a(\psi, \theta, \boldsymbol{T}, \hat{n}), \quad c \in \boldsymbol{A}, \tag{7}$$

where μ^ℓ is the local offset of the primitive, $\boldsymbol{R}$ and $\boldsymbol{S}$ are its rotation and scale, α is its opacity, c is its view-independent color and $\hat{n}$ is the surface normal of the corrected mesh. The final color combines c with an anisotropic spherical Gaussian appearance field Yang et al. (2024), which models view-dependent specular appearance from $\hat{n}$ and the viewing direction.

Basis-invariant position. Existing methods express the local attributes of a primitive in a frame built from the edges of its triangle and transform them to world space with that frame Qian et al. (2024);

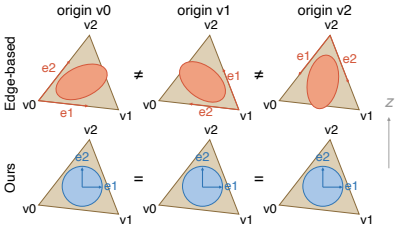


Figure 3: **The same primitive attributes under the edge basis of each vertex and J_w.**

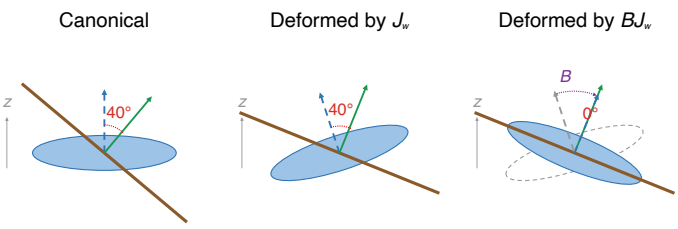


Figure 4: **Left**: canonical configuration. **Middle**: deformation by J_w misaligns the primitive normal with the surface normal. **Right**: B restores surface alignment.

Lee et al. (2026a). Because such a frame depends on the choice and ordering of the triangle vertices, identical local attributes can yield different global configurations (Fig. 3). We instead express the local offset in the canonical configuration, which is one fixed frame for every triangle, and transport it with the deformation gradient J_w that maps the canonical triangle to its corrected configuration. Since J_w is invariant to the choice of local basis, the resulting transform depends only on the surface deformation (Appendix D). The position of each primitive is therefore given by

$$\mu = J_w \, \mu^{\ell} + p, \tag{8}$$

where p is the surface point on the corrected mesh and μ^{ℓ} is the offset of the primitive expressed in the canonical configuration.

Surface-aligned covariance. Applying J_w alone does not guarantee surface alignment even when $R = I$, since the attributes are defined in the canonical configuration and the normal of an identity-rotation primitive is the normalized cross product of the first two columns of J_w (Fig. 4). We therefore define an alignment basis B as the minimal rotation that maps this normal onto the surface normal, and compose it with the deformation gradient in the transform of the local covariance $\Sigma^{\ell} = RSS^{\top}R^{\top}$:

$$\Sigma = \left(BJ_w\right) \Sigma^{\ell} \left(BJ_w\right)^{\top}. \tag{9}$$

With this construction, $R = I$ corresponds to a surface-aligned primitive in every configuration (Appendix E), while R represents only the learned deviation from the surface. Thus, both the primitive position and covariance consistently follow the deformation of the corrected surface.

3.4 Training Objectives

Geometry stage losses. The geometry stage (Sec. 3.2) optimizes the corrective network Φ using a proxy renderer composed of a static per-vertex texture a and global second-order spherical harmonics lighting ℓ Ramamoorthi & Hanrahan (2001). These auxiliary variables are used only to provide photometric supervision for the corrected geometry. We render the corrected mesh V_w with diffuse shading as $I = \mathcal{R}(V_w, a, \ell)$ and optimize

$$\mathcal{L}_{\text{geo}} = \mathcal{L}_1 + \lambda_{\text{wb}} \, \mathcal{L}_{\text{wb}} + \lambda_{\text{lap}} \, \mathcal{L}_{\text{lap}} + \lambda_{\text{d}} \, \mathcal{L}_{\text{d}} + \lambda_{\text{n}} \, \mathcal{L}_{\text{n}}, \tag{10}$$

where $\mathcal{L}_1$ is the photometric ℓ_1 loss between rendered and input images over the foreground region.

The wrinkle-band loss follows the principle of shading-based refinement Zollhöfer et al. (2015). Let $I_\ell = \mathcal{R}(V_w, 1, \ell)$ denote the shading image rendered with a uniform white texture, and let $\mathcal{H}(x) = x - G_\sigma * x$ denote a Gaussian high-pass operator, which we apply to log luminance. We match the high-frequency structure of the rendered shading to that of the input image using local normalized cross-correlation:

$$\mathcal{L}_{\text{wb}} = \frac{1}{|M|} \sum_{p \in M} \Big(1 - \text{NCC}_k\big(\mathcal{H}(\log I_\ell),\, \mathcal{H}(\log I^{\text{gt}})\big)_p\Big), \tag{11}$$

where NCC_k denotes zero-normalized cross-correlation over $k \times k$ patches and M is the rasterized face region. Because I_ℓ contains no texture and the gradient to ℓ is stopped in this term, its high-frequency variation can only be explained by the deforming surface. The resulting gradient therefore reaches ΔJ through the surface normals, encouraging wrinkle geometry whose spatial structure matches the high-frequency shading cues in the input. The normalized correlation further reduces sensitivity to local intensity gain and bias.

Table 1: **Quantitative comparison on NeRSemble** Kirschstein et al. (2023).

	Novel-View Synthesis			Novel Expression		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GaussianAvatars	31.07	0.936	0.125	25.76	0.910	0.134
SurFhead	29.83	0.930	0.149	25.96	0.913	0.149
TexAvatars	**35.56**	**0.954**	0.057	26.77	0.908	0.085
Ours	35.19	0.953	**0.055**	**27.26**	**0.914**	**0.079**

Table 2: **Normal accuracy** over the foreground (Full) and the face region (Face).

	Novel-View Synthesis		Novel Expression	
Method	Full	Face	Full	Face
GaussianAvatars	0.856	0.917	0.852	0.919
SurFhead	0.913	0.947	0.908	0.949
TexAvatars	0.857	0.924	0.855	0.925
Ours	**0.934**	**0.968**	**0.932**	**0.968**

Three regularizers constrain the wrinkle deformation. Using the uniform graph Laplacian $\boldsymbol{L}$ over the facial region, $\mathcal{L}_{\text{lap}} = \|\boldsymbol{L}(\boldsymbol{V}_w - \boldsymbol{V}_c)\|_2^2$ penalizes spatially irregular displacement, while $\mathcal{L}_{\text{n}} = \|\boldsymbol{L}N(\boldsymbol{V}_w)\|_2^2$ suppresses high-frequency variation in the vertex normals $N(\boldsymbol{V}_w)$. Finally, the displacement magnitude penalty $\mathcal{L}_{\text{d}} = \|\boldsymbol{V}_w - \boldsymbol{V}_c\|_2^2$ limits the magnitude of the correction and prevents low-frequency drift of the surface that is not penalized by the Laplacian term.

Appearance stage losses. The appearance stage optimizes the attribute decoders while keeping the geometry stage fixed. The objective is

$$\mathcal{L}_{\text{app}} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{vgg}}\,\mathcal{L}_{\text{vgg}} + \lambda_{\mu}\,\mathcal{L}_{\mu} + \lambda_{\text{s}}\,\mathcal{L}_{\text{s}} + \lambda_{\text{align}}\,\mathcal{L}_{\text{align}}, \tag{12}$$

where $\mathcal{L}_{\text{rgb}} = (1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\text{D-SSIM}}$ measures the photometric discrepancy between the rendered and input images, and $\mathcal{L}_{\text{vgg}}$ is a perceptual loss computed on VGG features. Following Qian et al. (2024), the regularizers $\mathcal{L}_{\mu} = \|\max(\mu^{\ell}, \epsilon_{\mu})\|_2$ and $\mathcal{L}_{\text{s}} = \|\max(\boldsymbol{S}, \epsilon_{\text{s}})\|_2$ constrain the primitive offsets and scales, respectively, keeping the Gaussians close to their associated triangles and spatially compact. Finally, $\mathcal{L}_{\text{align}}$ encourages the primitives to remain thin and tangent to the corrected surface by penalizing normal misalignment together with the smallest primitive scale. The optimization settings of both stages are given in Appendix G.

4 Experiments

Dataset and baselines. We evaluate on the NeRSemble dataset Kirschstein et al. (2023) and use the tracked meshes and the split of GaussianAvatars Qian et al. (2024), which trains on nine of the ten expression sequences and on 15 of the 16 cameras at a resolution of 550×802. The held-out camera gives the novel-view evaluation and the held-out sequence the novel-expression evaluation. Each subject also provides a free performance outside the ten prescribed sequences. We compare with GaussianAvatars Qian et al. (2024), SurFhead Lee et al. (2025) and TexAvatars Lee et al. (2026a), which we train with their official code and evaluate with the same pipeline. MeGA Wang et al. (2025a) does not reconstruct the torso and a fair comparison under the same metrics is therefore not possible. We exclude it from the tables and compare the recovered geometry with it qualitatively. Appendix J lists the subjects and the held-out sequences.

4.1 Quantitative Results

Tab. 1 reports the rendering quality. For novel-view synthesis, TexAvatars achieves higher PSNR and SSIM, while our method achieves lower LPIPS. For novel-expression synthesis, where an entire expression sequence is held out from training, our method achieves the best average performance on all three metrics. To quantify surface-normal accuracy, Tab. 2 reports the normal cosine similarity (NCS) between the rendered normals and the pseudo ground-truth normals of DAViD Saleh et al. (2025). These normals are used for evaluation only, and no 3D supervision is used to train any method. The normal of a 3D Gaussian is defined as its shortest principal axis, while SurFhead directly provides the normals of its surfels. Our method achieves the highest NCS in all four evaluation settings. SurFhead is the closest baseline and keeps its surfels aligned with the tracked mesh, whereas the primitives of GaussianAvatars and TexAvatars are not constrained to the surface. To directly evaluate the recovered geometry, we additionally compare the tracked and corrected meshes against multi-view stereo depth. As shown in Tab. 3, the corrected mesh reduces the mean depth error on both evaluation splits despite using no depth supervision. Since the stereo reconstruction does not resolve individual wrinkles, this comparison evaluates the recovered surface at the facial-shape scale rather than providing direct wrinkle-scale ground truth. Per-subject rendering and normal results are provided in Appendix L, and detailed mesh evaluations against multi-view stereo are provided in Appendix H.

Table 3: Mean mesh depth MAE (mm) against multi-view stereo.

	Novel View ↓	Novel Expression ↓
Tracked mesh	1.89	2.97
Corrected mesh	**1.17**	**2.46**

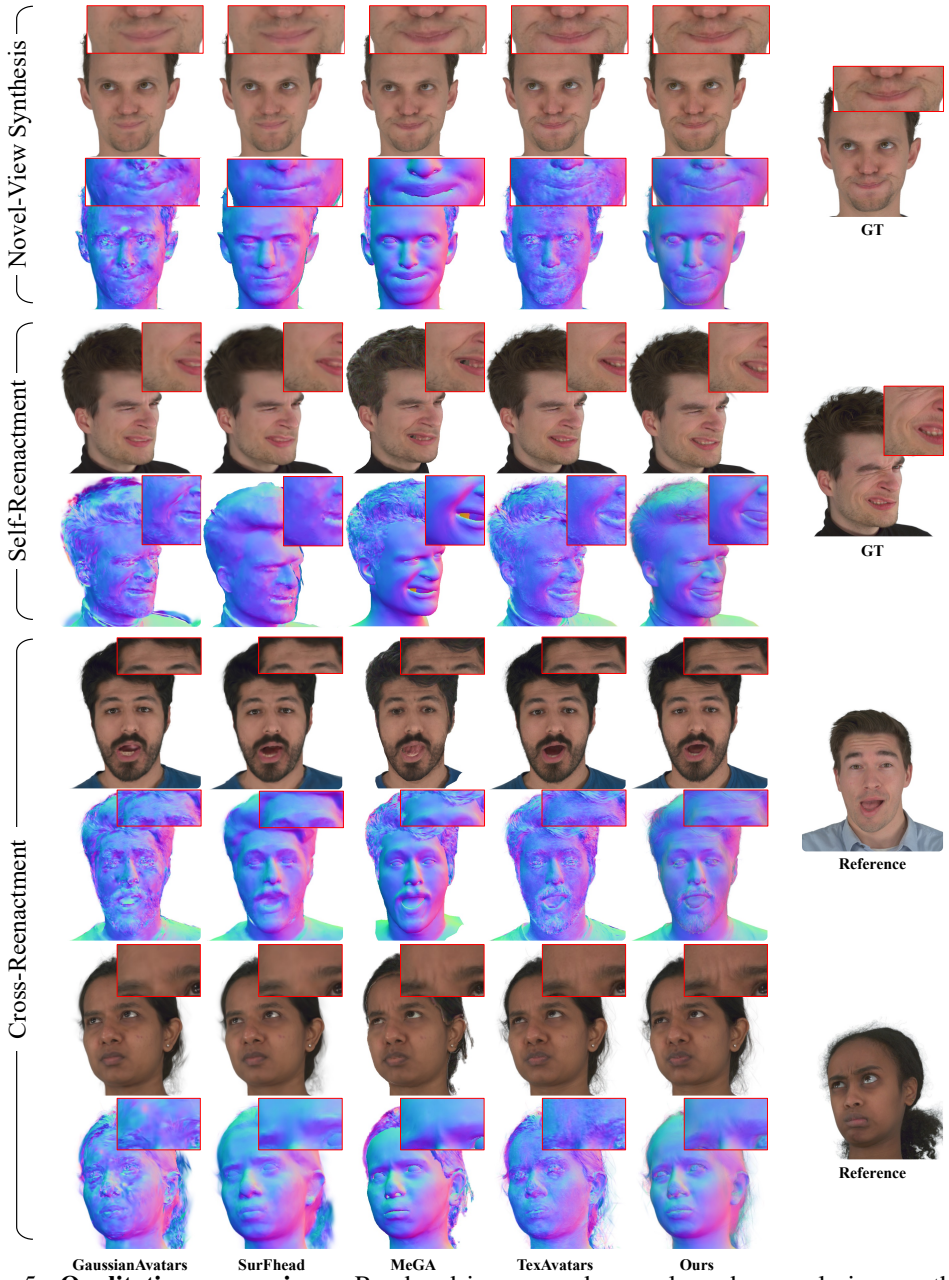


Figure 5: **Qualitative comparison.** Rendered images and normals under novel-view synthesis, self-reenactment and cross-reenactment. The last column shows the captured frame and for cross-reenactment the driving subject.

4.2 Qualitative Results

Fig. 5 compares the rendered images and normals of all methods under novel-view synthesis, self-reenactment, and cross-reenactment. The normals of the Gaussian-based methods are computed as in Tab. 2, while MeGA renders the normals of its facial mesh. Cross-reenactment drives each avatar with the free performance of another subject, which contains stronger expressions than the prescribed sequences. TexAvatars and MeGA reproduce visible wrinkles around the mouth and forehead, while our method exhibits finer wrinkle structure in the rendered normals. The normals of GaussianAvatars and TexAvatars are noisy, whereas SurFhead produces a smoother surface without the corresponding wrinkle structure. Although MeGA is supervised with multi-view stereo depth, this depth does not resolve individual wrinkles, and its reconstructed surface remains smooth in these regions. In contrast, our corrected geometry captures forehead and inter-brow wrinkles under cross-reenactment. Fig. 6 further compares the tracked and corrected meshes under wrinkle-forming



Figure 6: **Geometry recovered by our method.** Each column is a subject under an expression that forms wrinkles and the meshes are shaded with their normals.

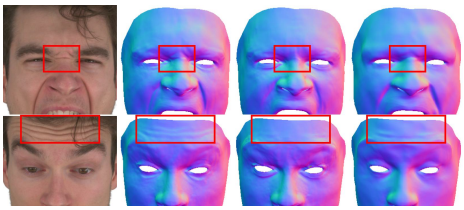


Figure 7: **Normals of ours without the strain map or the wrinkle-band loss.**

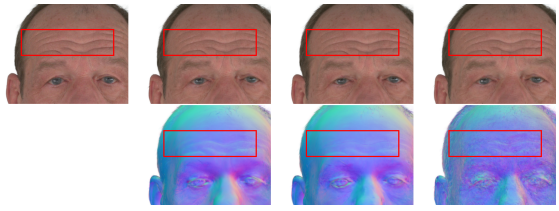


Figure 8: **Effect of the corrected mesh and the alignment term.**

expressions. While the tracked mesh remains smooth, the corrected mesh recovers wrinkle structures beside the mouth and around the eyes and nose that are visible in the reference images. Additional qualitative results are provided in Appendix K, and Appendix H compares our corrected mesh and the facial mesh of MeGA against multi-view stereo depth.

4.3 Ablation Study

Without the strain map $\boldsymbol{T}$, the wrinkles are largely absent, while removing the wrinkle-band loss $\mathcal{L}_{\text{wb}}$ suppresses fine wrinkle detail (Fig. 7); both variants also increase the stereo-depth error reported in Appendix H. In Tab. 4, removing the alignment basis $\boldsymbol{B}$ reduces both PSNR and NCS, whereas removing the alignment loss $\mathcal{L}_{\text{align}}$ leaves rendering quality nearly unchanged but reduces NCS by a similar margin. Removing the corrected mesh $\boldsymbol{V}_w$ and $\boldsymbol{T}$ from the decoders reduces PSNR but has little effect on NCS, as the pseudo ground-truth normals do not resolve wrinkle-scale normal variations. On the free-performance sequence in Fig. 8, the rendered images remain similar because the appearance stage can absorb the photometric residual, whereas the primitive normals follow the forehead wrinkles only when coupled to the corrected mesh and become noisy without the alignment loss.

Table 4: **Ablations on 104 subject under novel-view synthesis.** Each row removes one component. NCS is the normal cosine similarity over the full foreground.

	PSNR↑	SSIM↑	LPIPS↓	NCS↑
Ours	**34.46**	**0.946**	0.075	**0.912**
− corrected mesh $\boldsymbol{V}_w$	33.98	0.945	0.076	**0.912**
− $\boldsymbol{T}$ in decoders	34.26	**0.946**	0.075	**0.912**
− alignment basis $\boldsymbol{B}$	33.56	0.945	0.076	0.895
− alignment loss $\mathcal{L}_{\text{align}}$	34.45	**0.946**	**0.073**	0.895

5 Conclusion

We presented StrainAvatar, a head avatar that recovers expression-dependent skin deformation from multi-view RGB. We analytically derive strain from the tracked deformation for each expression and use a per-subject network to predict the corresponding corrective field. A Poisson solve integrates these correctives into a corrected mesh, with a zero corrective recovering the tracked mesh, such that learning is restricted to the identity-specific skin response. The Gaussian primitives are then transported by the deformation gradient of the corrected mesh and kept aligned with its surface. Experiments show rendering quality comparable to recent head avatars, improved surface-normal accuracy, and consistent reductions in mesh error relative to the tracked geometry, while qualitatively recovering expression-dependent wrinkle structure as explicit, animatable geometry.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods, supporting data analysis and interpreting results. We have not used generative AI tools for generating synthetic data, developing theoretical models, formulating mathematical claims or proofs, proposing hypotheses or designing the research methodology, and translation and dataset cleaning are not applicable to this work. Additionally, we used generative AI tools for editing the text for readability and consistency, creating and editing evaluation and visualization code, creating figures, searching and summarizing literature and formatting references. We have reviewed all AI-assisted work. All edited text was checked by the authors against the method and the measurements, AI-written evaluation code was verified by reproducing previously recorded results, and every reported number was measured with our own pipeline. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our method reconstructs a photorealistic and animatable avatar of a specific person and cross-reenactment renders that person under the expressions of another subject. This capability carries the risk of impersonation and fabricated video. The method requires a synchronized multi-view capture of the person with tracked meshes and trains one model per identity. It does not apply to a photograph or a video taken without the participation of that person and it does not transfer across identities. We train and evaluate on the NeRSemble dataset alone and record no new subjects. Its participants signed a GDPR-compliant consent agreement Kirschstein et al. (2023) and the faces shown in this paper appear under the terms of the dataset. We will release the code for research use only. The evaluation covers nine subjects and does not represent the diversity of human faces. Appendix H shows that the corrected mesh represents only the skin and the accuracy on faces outside these subjects remains to be verified.

REPRODUCIBILITY STATEMENT

The geometry stage and the appearance stage are described in Sec. 3.2 and Sec. 3.3. Appendix A gives the boundary conditions of the Poisson solve, Appendix C and Appendix F the architectures of the corrective network and the attribute decoders and Appendix G the loss weights and the training schedule. Appendix J lists the subjects, the sequences and the splits of the public NeRSemble dataset Kirschstein et al. (2023) together with the evaluation settings of every compared method, and Appendix L reports the per-subject results behind every average. We will release the code, the configuration files and the evaluation scripts.

REFERENCES

Noam Aigerman, Kunal Gupta, Vladimir G Kim, Siddhartha Chaudhuri, Jun Saito, and Thibault Groueix. Neural Jacobian fields: Learning intrinsic mappings of arbitrary meshes. *ACM Transactions on Graphics*, 41(4), 2022.

Shivangi Aneja, Sebastian Weiss, Irene Baeza, Prashanth Chandran, Gaspard Zoss, Matthias Niessner, and Derek Bradley. ScaffoldAvatar: High-fidelity Gaussian avatars with patch expressions. In *ACM SIGGRAPH 2025 Conference Papers*, pp. 1–11, 2025.

Shrisha Bharadwaj, Yufeng Zheng, Otmar Hilliges, Michael J Black, and Victoria Fernandez-Abrevaya. Flare: Fast learning of animatable and relightable mesh avatars. *arXiv preprint arXiv:2310.17519*, 2023.

Bernd Bickel, Manuel Lang, Mario Botsch, Miguel A Otaduy, and Markus Gross. Pose-space animation and transfer of facial details. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, pp. 57–66, 2008.

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. In *European Conference on Computer Vision*, pp. 459–476. Springer, 2024.

Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8498–8507. IEEE, 2024.

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. NPGA: Neural parametric Gaussian avatars. In *SIGGRAPH Asia 2024 conference papers*, pp. 1–11, 2024.

Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular RGB videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022.

Jingfan Guo, Jae Shin Yoon, Shunsuke Saito, Takaaki Shiratori, and Hyun Soo Park. High-fidelity modeling of generalizable wrinkle deformation. In *European Conference on Computer Vision*, pp. 429–445. Springer, 2024.

Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5322–5332. IEEE, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. NeRSemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023.

Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (TOG)*, 2020.

Jaeseong Lee, Taewoong Kang, Marcel Buehler, Min-Jung Kim, Sungwon Hwang, Junha Hyung, Hyojin Jang, and Jaegul Choo. SurFhead: Affine rig blending for geometrically accurate 2D Gaussian surfel head avatars. In *International Conference on Learning Representations (ICLR)*, 2025.

Jaeseong Lee, Junyeong Ahn, Taewoong Kang, and Jaegul Choo. TexAvatars: Hybrid texel-3D representations for stable rigging of photorealistic Gaussian head avatars. In *2026 International Conference on 3D Vision (3DV)*, pp. 1927–1944. IEEE, 2026a.

Seonghak Lee, Junhee Cho, Jisoo Park, Min-Gyu Park, Jongmin Lee, Ju Hong Yoon, and Junseok Kwon. Urhead: A unified uv-space representation for joint mesh–3dgs optimization in head avatars. In *European Conference on Computer Vision*, pp. 464–481. Springer, 2026b.

John P Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Frederic H Pighin, and Zhigang Deng. Practice and theory of blendshape facial models. *Eurographics State of the Art Reports*, 1(8):2, 2014.

Gengyan Li, Paulo Gotardo, Timo Bolkart, Stephan Garbin, Kripasindhu Sarkar, Abhimitra Meka, Alexandros Lattas, and Thabo Beeler. TeGA: Texture space Gaussian avatars for high-resolution dynamic head modeling. In *ACM SIGGRAPH 2025 Conference Papers*, pp. 1–12, 2025.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6):194:1–194:17, 2017.

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018.

Wan-Chun Ma, Andrew Jones, Jen-Yuan Chiang, Tim Hawkins, Sune Frederiksen, Pieter Peers, Marko Vukovic, Ming Ouhyoung, and Paul Debevec. Facial performance synthesis using deformation-driven polynomial displacement maps. *ACM Transactions on Graphics (TOG)*, 27 (5), 2008.

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic head avatars with rigged 3D Gaussians. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20299–20309. IEEE, 2024.

Dafei Qin, Jun Saito, Noam Aigerman, Thibault Groueix, and Taku Komura. Neural face rigging for animating and retargeting facial meshes in the wild. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023.

Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 497–500, 2001.

Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 130–141. IEEE, 2024.

Fatemeh Saleh, Sadegh Aliakbarian, Charlie Hewitt, Lohit Petikam, Xiao-Xian, Antonio Criminisi, Thomas J. Cashman, and Tadas Baltrušaitis. DAViD: Data-efficient and accurate vision models from synthetic data, 2025. URL `https://arxiv.org/abs/2507.15365`.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1606–1616. IEEE, 2024.

Eftychios Sifakis and Jernej Barbic. Fem simulation of 3d deformable solids: a practitioner's guide to theory, discretization and model reduction. In *Acm siggraph 2012 courses*, pp. 1–50. 2012.

Robert W Sumner and Jovan Popović. Deformation transfer for triangle meshes. *ACM Transactions on graphics (TOG)*, 23(3):399–405, 2004.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

Cong Wang, Di Kang, Heyi Sun, Shenhan Qian, Zixuan Wang, Linchao Bao, and Song-Hai Zhang. MeGA: Hybrid mesh-Gaussian head avatar for high-fidelity rendering and head editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26274–26284. IEEE, 2025a.

Rong Wang, Fabian Prada, Ziyan Wang, Zhongshi Jiang, Chengxiang Yin, Junxuan Li, Shunsuke Saito, Igor Santesteban, Javier Romero, Rohan Joshi, Hongdong Li, Jason Saragih, and Yaser Sheikh. FRESA: Feedforward reconstruction of personalized skinned avatars from few images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025b.

Yating Wang, Xuan Wang, Ran Yi, Yanbo Fan, Jichen Hu, Jingcheng Zhu, and Lizhuang Ma. 3D Gaussian head avatars with expressive dynamic appearances by compact tensorial representations. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21117–21126. IEEE, 2025c.

Changyeon Won, Min-Gyu Park, Seonghwan Park, Ju Hong Yoon, and Hae-Gon Jeon. Jacobiana-vatar: Temporally consistent semi-rigid avatar reconstruction from a monocular video. In *European Conference on Computer Vision*, pp. 641–659. Springer, 2026.

Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. FlashAvatar: High-fidelity head avatar with efficient Gaussian embedding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1802–1812. IEEE, 2024.

Jiahao Yang, Xiaohang Yang, Qing Wang, Yilan Dong, Gregory Slabaugh, and Shanxin Yuan. Self-learning expression deformations for data-efficient gaussian avatars. *arXiv preprint arXiv:2606.05912*, 2026.

Ziyi Yang, Xinyu Gao, Yangtian Sun, Yihua Huang, Xiaoyang Lyu, Wen Zhou, Shaohui Jiao, Xiaojuan Qi, and Xiaogang Jin. Spec-Gaussian: Anisotropic view-dependent appearance for 3D Gaussian splatting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Hongwen Zhang, Siyou Lin, Ruizhi Shao, Yuxiang Zhang, Zerong Zheng, Han Huang, Yandong Guo, and Yebin Liu. CloSET: Modeling clothed humans on continuous surface with explicit template decomposition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 501–511. IEEE, 2023.

Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13535–13545. IEEE, 2022.

Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. PointAvatar: Deformable point-based head avatars from videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21057–21067. IEEE, 2023.

Michael Zollhöfer, Angela Dai, Matthias Innmann, Chenglei Wu, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics (ToG)*, 34(4):1–14, 2015.

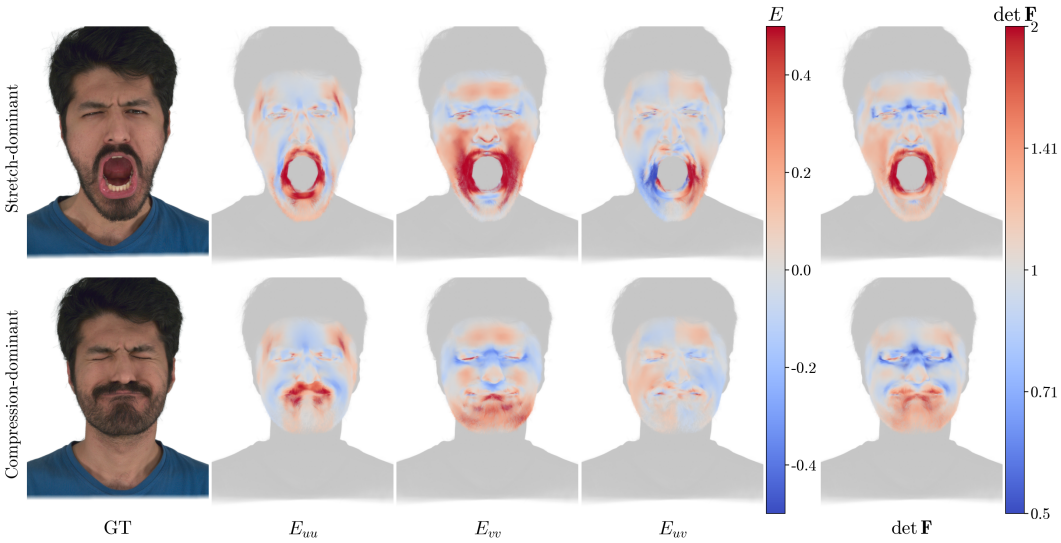


Figure 9: **Strain field.** Both rows share one color scale, with the strain clipped to $[-0.5, 0.5]$ and the area change to $[0.5, 2]$ on a logarithmic axis. Gray marks the head outside the face region, where the correction is not defined.

A Boundary Conditions of the Poisson Solve

The solve of Eq. 6 requires boundary conditions for two reasons. First, the Poisson system determines vertex positions only up to a global translation, so an unconstrained solve is not well posed. NJF Aigerman et al. (2022) resolves the ambiguity by fixing a single vertex of the output. In our setting the face region is part of a larger head mesh, so the whole boundary is constrained instead. Second, anchoring the boundary keeps the corrected face region attached to the surrounding tracked head without seams or drift under every expression.

The boundary variables stay in the system and the constraint is implemented as a quadratic penalty. With the diagonal matrix $\boldsymbol{\Omega}$ of triangle areas and the diagonal indicator $\boldsymbol{D}_\partial$ of the boundary vertices, the corrected mesh is the solution of

$$\left(\nabla^\top \boldsymbol{\Omega} \nabla + \lambda_\partial \boldsymbol{D}_\partial\right) \boldsymbol{V}_w = \nabla^\top \boldsymbol{\Omega} \tilde{\boldsymbol{J}} + \lambda_\partial \boldsymbol{D}_\partial \boldsymbol{V}_c, \tag{13}$$

where $\tilde{\boldsymbol{J}}$ stacks the corrected Jacobians of Eq. 5 and $\nabla^\top \boldsymbol{\Omega} \nabla$ is the cotangent Laplacian of the face region. The penalty weight is $\lambda_\partial = 10^8$ times the median diagonal entry of that Laplacian, so the boundary vertices coincide with the tracked mesh up to numerical precision while the system remains symmetric positive definite. The augmented system is prefactorized with a sparse Cholesky decomposition.

B Strain Field

Fig. 9 shows the three components of the strain signal s^f of Eq. 3 for a stretch-dominant and a compression-dominant expression of the same subject. Their rasterization over the face UV map is the strain map $\boldsymbol{T}$ that provides the input to Φ. We additionally visualize the area change

$$\det \boldsymbol{F}^f = \sqrt{\det\left(\boldsymbol{I} + 2\boldsymbol{E}^f\right)}, \tag{14}$$

which follows from Eq. 2 and satisfies $\det \boldsymbol{F}^f = 1$ where the triangle keeps its canonical area. It is not an input of Φ. Each quantity is shown by coloring the primitives of Sec. 3.3 with the value of their triangle and rendering the avatar from the held-out camera.

The strain signal is derived analytically from the tracked mesh and is defined under every expression. The open mouth of the top row stretches the skin around the lips where $\det \boldsymbol{F}^f > 1$. The frown of the bottom row compresses the region between the brows and around the eyes where $\det \boldsymbol{F}^f < 1$. Skin buckles into folds under in-plane compression Bickel et al. (2008), so the compressive regions of the signal mark where the learned correction of Eq. 5 applies.

The sign of E^f_{uv} reverses between the two sides of the face because the u axis of the UV map is mirrored across the midline. This is a property of the map and not of the deformation.

C Corrective Network

The network Φ of Sec. 3.2 is a four-level UNet over the UV map with 16, 32, 64 and 128 channels. Every level applies two 3×3 convolutions with SiLU activations and no normalization. A strided convolution halves the resolution at each level after the first. The expansion path restores the resolution by bilinear upsampling and concatenates the encoder feature of the matching resolution. Each encoder level adds a bias map of a single channel that is broadcast over the channels. The final layer is a 3×3 convolution to the nine entries of $\Delta \boldsymbol{J}^f$ and is zero-initialized. Its output is passed through $s \tanh(\cdot/s)$ with $s = 2$, which acts as the identity on small correctives and bounds every entry of $\Delta \boldsymbol{J}^f$. The convolutions hold 0.54M parameters and the bias maps 0.35M.

D Basis Invariance of $\boldsymbol{J}_w$

Existing methods Qian et al. (2024); Lee et al. (2026a) express the attributes of a primitive in a basis built from the edge vectors of its triangle and transform them with the frame matrix of that basis alone. Such a basis is tied to the vertex ordering, so identical attributes at neighboring triangles are transformed into different global configurations and have no consistent geometric meaning across the face. The appearance stage of Sec. 3.3 instead transforms each primitive by the deformation gradient $\boldsymbol{J}_w = \boldsymbol{M}_f \bar{\boldsymbol{M}}_f^{-1} \in \mathbb{R}^{3\times3}$, where $\bar{\boldsymbol{M}}_f, \boldsymbol{M}_f \in \mathbb{R}^{3\times3}$ are the frame matrices of triangle f on the canonical and on the corrected mesh. Each frame matrix maps a local basis of the triangle to world space and the local basis itself is a convention.

A change of the local basis is a right multiplication of both frame matrices by a common invertible matrix $\boldsymbol{P} \in \mathbb{R}^{3\times3}$. A transform that composes the local attributes with a single frame matrix retains the factor $\boldsymbol{P}$, so the identity-rotation primitive changes with the vertex ordering (Fig. 3). The deformation gradient is unaffected,

$$(\boldsymbol{M}_f \boldsymbol{P})(\bar{\boldsymbol{M}}_f \boldsymbol{P})^{-1} = \boldsymbol{M}_f \boldsymbol{P} \boldsymbol{P}^{-1} \bar{\boldsymbol{M}}_f^{-1} = \boldsymbol{M}_f \bar{\boldsymbol{M}}_f^{-1} = \boldsymbol{J}_w, \tag{15}$$

since the same $\boldsymbol{P}$ enters both frames. The attributes can therefore be expressed directly in the world coordinates of the canonical mesh, which is the fixed frame of Sec. 3.3, and the primitives depend on the deformation of the mesh alone.

E Surface Alignment Across Configurations

The alignment basis $\boldsymbol{B}$ of Sec. 3.3 re-expresses the local basis of a primitive rather than initializing its rotation. Let

$$\hat{j} = \frac{\boldsymbol{J}_w e_1 \times \boldsymbol{J}_w e_2}{\|\boldsymbol{J}_w e_1 \times \boldsymbol{J}_w e_2\|} \tag{16}$$

denote the normal of an identity-rotation primitive, where e_1 and e_2 are the first two axes of the local basis. An alternative to the alignment basis is to initialize the rotation so that the primitive is aligned with the surface in one configuration. Both $\boldsymbol{J}_w$ and the surface normal $\hat{n}$ change with the expression, so a rotation stored for one configuration is no longer surface-aligned in the next and the network would have to compensate for the residual misalignment at every frame. A prior that regularizes the rotation toward the identity would then oppose the alignment.

The alignment basis is instead defined per configuration. It is the minimal rotation from $\hat{j}$ to $\hat{n}$ of the current configuration, so

$$\boldsymbol{B}\, \hat{j} = \hat{n} \tag{17}$$

holds under every expression and $\boldsymbol{B}$ has no learnable parameters. The rotation of Eq. 9 therefore describes only the deviation from the surface-aligned state.

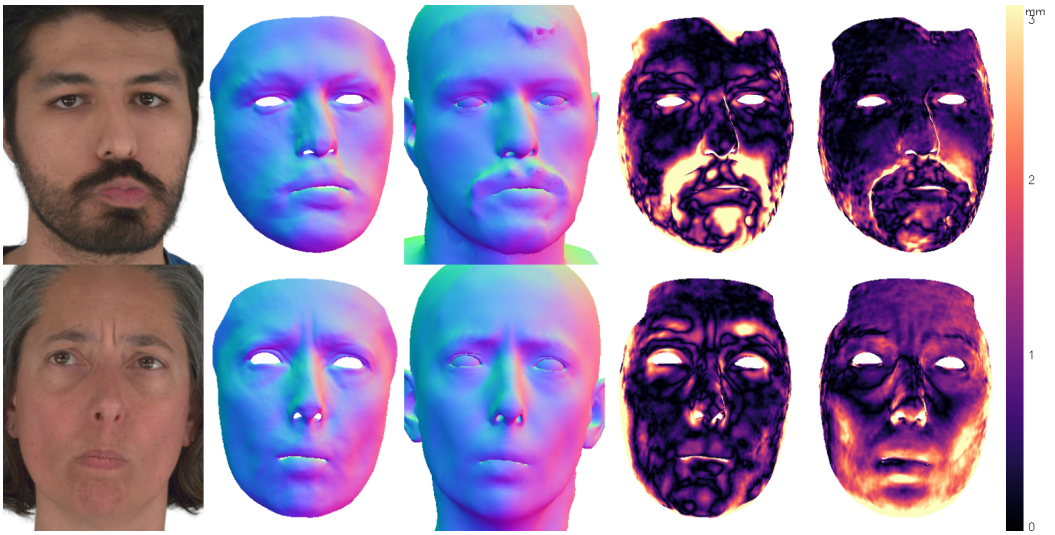


Figure 10: **Corrected mesh and the facial mesh of MeGA against multi-view stereo.** Both meshes are shaded with their normals and the error columns show the distance to the stereo depth over the same mask.

F Attribute Decoders

Both decoders of Sec. 3.3 follow the architecture of TexAvatars Lee et al. (2026a), which in turn follows RGCA Saito et al. (2024). A linear layer maps the expression parameters ψ and the pose parameters θ of the tracked frame to a feature map of 256 channels at a resolution of 8×8. Transposed convolutions expand this map to the UV map resolution. The geometry decoder $\mathcal{D}_g$ outputs the eleven channels of $\boldsymbol{G}$, namely the local offset μ^ℓ, the rotation $\boldsymbol{R}$ as a quaternion, the scale $\boldsymbol{S}$ as a logarithm and the opacity α as a logit. The appearance decoder $\mathcal{D}_a$ outputs the three channels of $\boldsymbol{A}$, namely the view-independent color c. Its output bias is initialized to 0.5. A darker initial color can collapse to a transparent solution on the white background of the capture.

We keep the architecture and change the conditioning. The expression encoder receives ψ and θ without the EMOPortraits embedding Drobyshev et al. (2024) of TexAvatars, so the decoders are driven by the tracked parameters alone. The strain map $\boldsymbol{T}$ enters as a separate spatial condition. $\mathcal{D}_g$ takes $\boldsymbol{T}$ at the 128^2 and 256^2 stages. $\mathcal{D}_a$ takes $\boldsymbol{T}$ with the surface normal $\hat{n}$ at the 64^2 and 128^2 stages. The anisotropic spherical Gaussian appearance field carries a zero-initialized feature map of 24 channels over the UV map. A linear layer maps it to the parameters of 32 lobes Yang et al. (2024).

G Implementation Details

All components are implemented in PyTorch with nvdiffrast Laine et al. (2020) for differentiable mesh rasterization. The face region is subdivided twice by midpoint subdivision so that the wrinkles are resolved and all UV maps use a resolution of $R = 512$. The resulting face submesh has 27,608 vertices and 54,528 triangles. The primitives are bound to the valid pixels of the UV maps, so their number is fixed and no densification or pruning is used. Both stages are optimized with Adam Kingma & Ba (2015).

The geometry stage runs for 200k iterations with $\lambda_{\text{wb}} = 0.3$, $\lambda_{\text{lap}} = 10^6$, $\lambda_d = 3 \times 10^3$ and $\lambda_{\text{n}} = 10$. It uses a learning rate of 10^{-3} for Φ and 10^{-2} for the texture and the lighting and every iteration takes all training views of one timestep. The wrinkle-band correlation is evaluated over 15×15 patches. The tracked FLAME parameters stay fixed in this stage.

The appearance stage runs for 600k iterations with $\lambda = 0.2$ for the D-SSIM term and the perceptual term enabled after 300k iterations. It uses a learning rate of 6×10^{-4} for both decoders and finetunes the tracked FLAME parameters with learning rates 10^{-5} for the head rotation and the neck, jaw and

Table 5: **Depth error against multi-view stereo.** MAE of the rasterized depth in mm on the held-out camera with the lower of our two meshes in bold. †Trained with this stereo depth.

	Novel-View Synthesis			Novel Expression		
Subject	Tracked↓	Corrected↓	MeGA† ↓	Tracked↓	Corrected↓	MeGA† ↓
074	2.23	**1.39**	0.89	3.81	**3.12**	2.84
104	2.26	**1.69**	1.29	3.75	**3.33**	3.05
218	1.46	**1.03**	0.89	3.01	**2.79**	2.55
253	1.62	**1.05**	0.86	1.44	**1.09**	0.88
264	2.53	**1.32**	1.06	3.87	**2.81**	2.49
302	2.07	**1.14**	1.06	3.18	**2.45**	2.42
304	1.51	**0.89**	1.38	2.38	**2.08**	2.27
460	1.42	**0.86**	0.90	2.36	**2.04**	2.47
Average	1.89	**1.17**	1.04	2.97	**2.46**	2.37

eye poses and 10^{-6} for the translation. The expression parameters are not optimized, since the geometry stage has already fit the facial deformation and a free expression absorbs the photometric residual. Training takes about 11 hours per stage on a single NVIDIA RTX 6000 Ada GPU and excludes the periodic evaluation.

H Mesh Accuracy Against Multi-View Stereo

The pseudo ground truth of Tab. 2 comes from a learned estimator while the multi-view stereo depth of the dataset is reconstructed from the captured images without a learned prior. We rasterize the tracked mesh and the corrected mesh from the held-out camera and compare their depth with the stereo depth. Tab. 5 reports the mean absolute error (MAE) between the two for every subject that provides stereo depth and follows the Geo. MAE of MeGA Wang et al. (2025a). The table also lists the facial mesh of MeGA that is refined under the supervision of this stereo depth and serves as the reference for the error that a mesh fitted to the depth reaches. The two methods fit their mesh under different masks and we measure both over the hair-free intersection of the two masks. One subject provides no stereo depth and we exclude from all columns the frames with a failed stereo reconstruction and the frames without a parsing map.

Our corrected mesh uses no depth supervision and reduces the error of the tracked mesh on every subject and on both splits. The corrective moves the surface toward an independent measurement of the same face, providing evidence that the recovered deformation reflects geometric correction rather than solely an appearance artifact. On two subjects the corrected mesh is below the reference on both splits. Fig. 10 shows the two subjects at the ends of the novel-view columns of Tab. 5. The subject in the top row wears a beard and the stereo depth follows the outer surface of the hair that the depth-supervised mesh of MeGA fits while our corrected mesh stays on the skin beneath. The subject in the bottom row has no facial hair and the corrected mesh is closer to the stereo depth than the mesh of MeGA across the face. The gap in the top row therefore measures the hair volume that our mesh does not represent and not the accuracy of the skin surface. The larger error on the novel expression originates in the tracked mesh and our local corrective inherits it by construction.

Fig. 11 shows that the tracked mesh deviates over the cheeks, the chin and the brow. A linear expression basis represents these regions with a smooth surface and the corrective recovers their wrinkles. Multi-view stereo resolves the coarse shape of the face and not the depth of a single wrinkle, so the comparison validates the recovered surface at the scale of the face.

I Illumination Editing

The shading function of GaussianShader Jiang et al. (2024) evaluates the illumination with the surface normals, so it shows qualitatively whether the geometry recovered by the first stage supports shading. We adopt it together with its learnable environment cube map. Both are designed for a static scene and we apply them to a dynamic sequence.

We then replace the environment map with a moving spotlight and re-render the sequence from the held-out camera (Fig. 12). No ground truth exists under the new illumination, so the result is qualitative. The diffuse color of a primitive is a free parameter, so the captured illumination remains in it and the new illumination adds shading and highlights on top. The shading model has no visibility term, so cast shadows are not reproduced.



Figure 11: **Mesh accuracy against multi-view stereo.** The meshes are shaded with their normals and the error columns show the distance to the stereo depth.

Table 6: **Sequence held out for the novel-expression evaluation**.

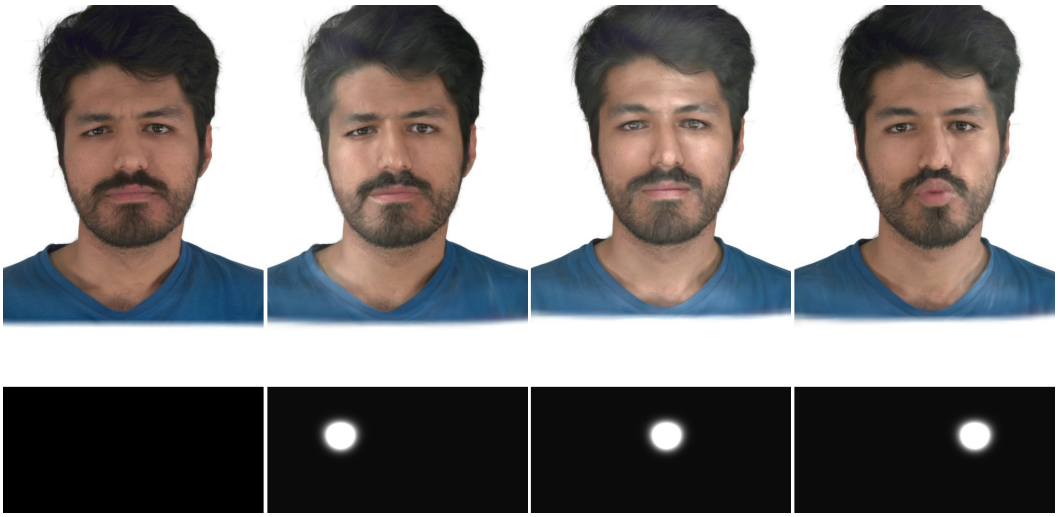
Subject	074	104	218	253	264	302	304	306	460
Sequence	EMO-4	EXP-2	EXP-9	EMO-4	EXP-9	EMO-2	EXP-2	EXP-2	EMO-3

J Evaluation Configuration

The nine subjects are 074, 104, 218, 253, 264, 302, 304, 306 and 460. Each provides ten prescribed sequences, namely EMO-1 to EMO-4 and EXP-2 to EXP-5, EXP-8 and EXP-9, together with a free performance. Tab. 6 lists the sequence held out from training for every subject and follows the choice of GaussianAvatars Qian et al. (2024).

K Additional Qualitative Results

Fig. 13 and Fig. 14 extend the comparison of Fig. 5 to further subjects.



Captured illumination Moving spotlight

Figure 12: **Illumination editing.** The environment map of each column is shown below the rendering.

Table 7: **Per-subject rendering quality on NeRSemble** Kirschstein et al. (2023).

		Novel-View Synthesis			Novel Expression		
Subject	Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
074 (EMO-4)	GaussianAvatars	29.76	0.921	0.180	**25.89**	0.894	0.172
	SurFhead	28.79	0.916	0.213	25.79	**0.896**	0.195
	TexAvatars	**36.26**	**0.947**	0.075	25.12	0.880	0.114
	Ours	35.95	0.946	**0.071**	25.00	0.884	**0.108**
104 (EXP-2)	GaussianAvatars	26.12	0.914	0.180	**24.11**	0.906	0.165
	SurFhead	25.02	0.908	0.210	23.80	**0.908**	0.186
	TexAvatars	**34.70**	**0.947**	0.080	22.88	0.894	0.116
	Ours	34.46	0.946	**0.075**	23.13	0.897	**0.112**
218 (EXP-9)	GaussianAvatars	28.99	**0.933**	0.097	25.89	0.914	0.113
	SurFhead	28.61	0.931	0.114	26.45	**0.922**	0.120
	TexAvatars	29.00	0.931	**0.051**	26.35	0.912	0.076
	Ours	**29.26**	0.929	0.053	**26.84**	0.921	**0.071**
253 (EMO-4)	GaussianAvatars	33.78	0.959	0.093	24.14	0.914	0.128
	SurFhead	31.95	0.952	0.112	24.06	0.914	0.139
	TexAvatars	**40.18**	**0.978**	**0.033**	26.77	0.922	0.073
	Ours	39.33	0.976	0.037	**29.28**	**0.939**	**0.058**
264 (EXP-9)	GaussianAvatars	30.50	0.933	0.139	29.15	**0.932**	0.128
	SurFhead	29.84	0.929	0.167	29.15	**0.932**	0.147
	TexAvatars	**33.92**	**0.946**	0.076	**29.19**	0.924	0.085
	Ours	33.51	**0.946**	**0.070**	29.08	0.924	**0.081**
302 (EMO-2)	GaussianAvatars	31.51	0.939	0.112	25.25	0.896	0.131
	SurFhead	30.02	0.930	0.131	25.69	0.898	0.142
	TexAvatars	34.84	0.958	0.054	26.13	0.900	0.079
	Ours	**34.84**	**0.959**	**0.051**	**26.60**	**0.907**	**0.073**
304 (EXP-2)	GaussianAvatars	28.49	0.884	0.157	25.07	0.853	0.165
	SurFhead	27.38	0.871	0.191	**25.12**	**0.861**	0.188
	TexAvatars	**33.38**	**0.925**	0.070	24.53	0.839	0.119
	Ours	33.21	**0.925**	**0.064**	24.55	0.842	**0.113**
306 (EXP-2)	GaussianAvatars	34.17	0.969	0.083	29.75	0.954	0.084
	SurFhead	32.99	0.965	0.097	31.03	0.957	0.091
	TexAvatars	**38.13**	**0.978**	0.036	**31.99**	**0.961**	**0.041**
	Ours	37.50	0.977	**0.035**	31.17	0.958	0.042
460 (EMO-3)	GaussianAvatars	36.29	0.971	0.082	22.57	0.925	0.119
	SurFhead	33.90	0.965	0.105	22.53	0.926	0.134
	TexAvatars	**39.60**	**0.977**	**0.041**	27.93	0.943	0.066
	Ours	38.69	0.975	0.042	**29.74**	**0.952**	**0.057**
Average	GaussianAvatars	31.07	0.936	0.125	25.76	0.910	0.134
	SurFhead	29.83	0.930	0.149	25.96	0.913	0.149
	TexAvatars	**35.56**	**0.954**	0.057	26.77	0.908	0.085
	Ours	35.19	0.953	**0.055**	**27.26**	**0.914**	**0.079**

L PER-SUBJECT RESULTS

Tab. 7 and Tab. 8 report the averages of Tab. 1 and Tab. 2 for every subject. Every entry is rendered and measured by us with the same pipeline. The sequence held out for the novel-expression evaluation differs across subjects and is given next to the subject identifier.

Table 8: **Per-subject normal accuracy against the pseudo ground truth of DAViD** Saleh et al. (2025).

		Novel-View Synthesis		Novel Expression	
Subject	Method	Full↑	Face↑	Full↑	Face↑
074 (EMO-4)	GaussianAvatars	0.8872	0.9071	0.8785	0.9065
	SurFhead	0.9179	0.9499	0.9046	0.9458
	TexAvatars	0.8699	0.9037	0.8574	0.8999
	Ours	**0.9422**	**0.9555**	**0.9358**	**0.9581**
104 (EXP-2)	GaussianAvatars	0.8415	0.9324	0.8367	0.9365
	SurFhead	0.9096	0.9503	0.9027	0.9440
	TexAvatars	0.8499	0.9343	0.8497	0.9339
	Ours	**0.9121**	**0.9704**	**0.9091**	**0.9700**
218 (EXP-9)	GaussianAvatars	0.8649	0.8842	0.8576	0.8861
	SurFhead	0.9115	0.9561	0.9018	0.9597
	TexAvatars	0.8915	0.9027	0.8818	0.9010
	Ours	**0.9461**	**0.9640**	**0.9398**	**0.9627**
253 (EMO-4)	GaussianAvatars	0.8542	0.9030	0.8451	0.9075
	SurFhead	0.8785	0.9493	0.8738	0.9527
	TexAvatars	0.8543	0.9377	0.8451	0.9397
	Ours	**0.9296**	**0.9696**	**0.9292**	**0.9714**
264 (EXP-9)	GaussianAvatars	0.9105	0.9487	0.9106	0.9516
	SurFhead	0.9540	0.9629	0.9529	0.9658
	TexAvatars	0.9069	0.9399	0.9052	0.9429
	Ours	**0.9586**	**0.9745**	**0.9559**	**0.9748**
302 (EMO-2)	GaussianAvatars	0.8559	0.9289	0.8609	0.9305
	SurFhead	0.9176	0.9220	0.9243	0.9325
	TexAvatars	0.8556	0.9106	0.8610	0.9123
	Ours	**0.9286**	**0.9596**	**0.9352**	**0.9603**
304 (EXP-2)	GaussianAvatars	0.7802	0.9089	0.7761	0.9047
	SurFhead	0.8750	0.9342	0.8679	0.9325
	TexAvatars	0.7799	0.9296	0.7809	0.9277
	Ours	**0.8930**	**0.9686**	**0.8920**	**0.9665**
306 (EXP-2)	GaussianAvatars	0.8776	0.9208	0.8697	0.9227
	SurFhead	0.9294	0.9570	0.9293	0.9620
	TexAvatars	0.8671	0.9369	0.8585	0.9366
	Ours	**0.9443**	**0.9759**	**0.9370**	**0.9760**
460 (EMO-3)	GaussianAvatars	0.8291	0.9171	0.8296	0.9238
	SurFhead	0.9235	0.9408	0.9192	0.9456
	TexAvatars	0.8422	0.9244	0.8566	0.9327
	Ours	**0.9561**	**0.9724**	**0.9571**	**0.9735**
Average	GaussianAvatars	0.8557	0.9168	0.8516	0.9189
	SurFhead	0.9130	0.9469	0.9085	0.9490
	TexAvatars	0.8575	0.9244	0.8551	0.9252
	Ours	**0.9345**	**0.9678**	**0.9323**	**0.9681**

M LIMITATIONS

The corrective is defined on the face region and its boundary is anchored to the tracked mesh. The correction is therefore local and inherits the error of the tracker. An expression that the tracked mesh misses produces no strain and no wrinkle and a global misalignment of the head passes into the corrected mesh. The corrected mesh represents only the skin and facial hair is carried by the primitives alone. Appendix I shades the geometry under a new illumination as a qualitative demonstration and the appearance is learned under the captured illumination and not separated from it. No ground truth exists at the scale of a wrinkle. The pseudo ground truth normals and the stereo depth resolve the coarse shape of the face and the evidence at the scale of a wrinkle is qualitative. The method trains one model per identity from a multi-view capture and does not transfer to another identity or to monocular video.

N ETHICAL CONSIDERATIONS

Our method reconstructs a photorealistic and animatable avatar of a specific person and cross-reenactment renders that person under the expressions of another subject. This capability carries the risk of impersonation and fabricated video. The method requires a synchronized multi-view capture of the person with tracked meshes and trains one model per identity. It does not apply to a photograph or a video taken without the participation of that person and it does not transfer across identities. We train and evaluate on the NeRSemble dataset alone and record no new subjects. Its participants signed a GDPR-compliant consent agreement Kirschstein et al. (2023) and the faces shown in this paper appear under the terms of the dataset. We will release the code for research use only. The evaluation covers nine subjects and does not represent the diversity of human faces. Appendix H shows that the corrected mesh represents only the skin and the accuracy on faces outside these subjects remains to be verified.

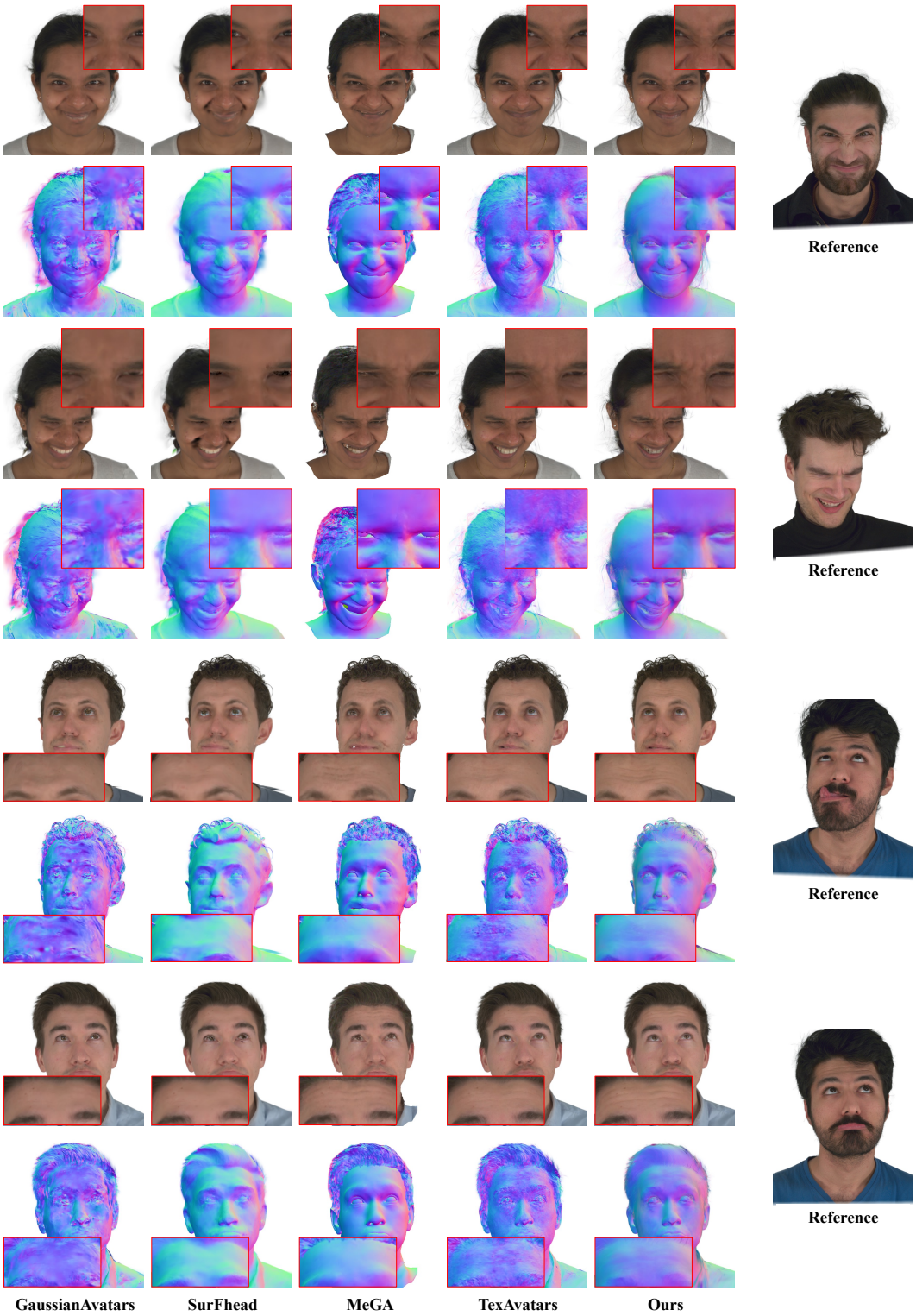


Figure 13: **Additional qualitative comparison.** Rendered images and normals of all methods on further subjects.

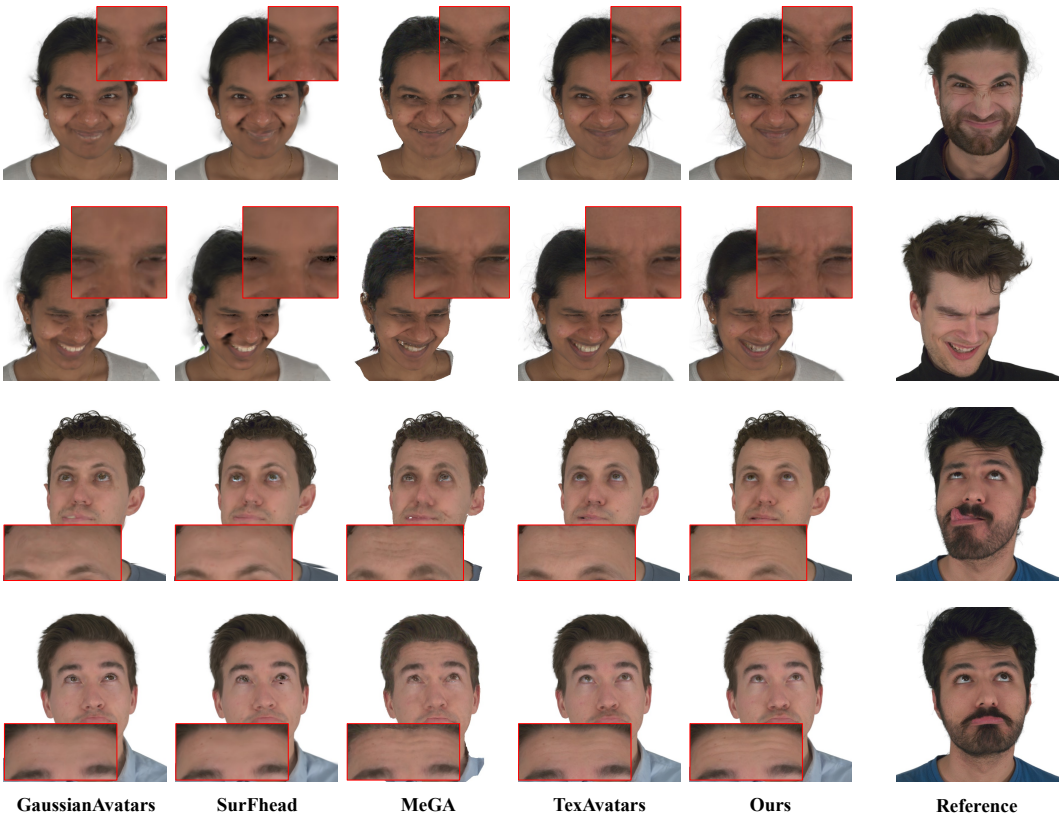


Figure 14: **Additional qualitative comparison.** Rendered images and normals of all methods on further subjects.