Feed-forward Gaussian Registration for Head Avatar Creation and Editing

Malte Prinzler[1,2*] Paulo Gotardo[2] Siyu Tang[1] Timo Bolkart[2]
[1]ETH Zürich [2]Google

Figure 1. Given calibrated multi-view images as input, MATCH infers static Gaussian splat textures in 0.5 seconds. The resulting Gaussians are in dense semantic correspondence across subjects and expressions. This enables diverse downstream applications such as fast head avatar creation, interpolation, semantic editing, and expression transfer. For visualization, we show 6 of the 12 input images, display predicted Gaussians for three separate frames, and apply a checkerboard semantic texture to highlight the dense correspondence.

Abstract

We present MATCH (Multi-view Avatars from Topologically Corresponding Heads), a multi-view Gaussian registration method for high-quality head avatar creation and editing. State-of-the-art multi-view head avatars require time-consuming head tracking followed by expensive avatar optimization, resulting in creation times exceeding one day. In contrast, MATCH directly predicts Gaussian splats in correspondence from calibrated multi-view images in 0.5 seconds per frame. While the learned intra-subject correspondence across frames allows us to quickly build personalized head avatars, correspondence across subjects enables expression transfer, optimization-free tracking, semantic editing, and interpolation. We establish these correspondences with a transformer that predicts textures of Gaussian splats. To this end, we introduce a novel attention block, in which each UV map token attends exclusively to image tokens depicting its corresponding mesh region. MATCH outperforms existing methods for novel-view synthesis, geometry registration, and head avatar generation, the latter being 10× faster than the qualitatively closest baseline. The code and model weights are available on the project website.

*Work done while Malte Prinzler was an intern at Google.

1. Introduction

The growing demand for realistic digital humans in telepresence, film, and gaming has intensified the need for scalable methods that can rapidly create controllable, photorealistic head avatars. Obtaining state-of-the-art head avatars from multi-view in-studio captures typically relies on a two-stage pipeline. This involves first establishing cross-view and temporal correspondence via mesh-based tracking before optimizing surface-attached primitives, such as Gaussian splats [22, 25], to model appearance [18, 53, 70]. Despite its high-fidelity results, this two-stage process is a major computational bottleneck, as creating a single personalized avatar typically requires hours, or even days, of optimization. Consequently, scaling to a large number of subjects becomes prohibitively expensive.

Our method MATCH overcomes these limitations by directly predicting Gaussian splats in dense semantic correspondence from multi-view images in 0.5 seconds per frame. In this context, dense semantic correspondence means that all output Gaussian textures share a fixed topology in which a specific Gaussian consistently represents the same semantic region (e.g., the nose tip), regardless of identity or expression. Crucially, this correspondence enables downstream applications such as building lightweight

head avatars in the form of linear Gaussian Eigen Models (GEM) [70], or performing direct expression transfer and semantic editing (see Figure 1). By bypassing both time-consuming mesh tracking (10.7 hours) and animatable Gaussians optimization (27.7 hours), our approach reduces the total GEM avatar creation time by a factor of 10.

MATCH employs a transformer architecture inspired by Large Reconstruction Models (LRM) [56, 65, 67]. The UV texture map and input images are tokenized and processed by a sequence of transformer blocks. However, naïvely attending to all image and UV tokens [21, 60, 61, 63] is computationally expensive and leads to poor generalization on unseen subjects [29]. Instead, we introduce a novel registration-guided attention block. It restricts each UV token to attend only to image tokens displaying the relevant head region, which reduces compute complexity while simultaneously improving synthesis quality.

In summary, MATCH represents heads with Gaussian textures in dense semantic correspondence that are regressed directly from multi-view input images using a novel registration-guided attention mechanism. These textures can subsequently be processed into lightweight GEM [70] avatars, drivable by monocular videos. Finally, we show that dense Gaussian correspondence enables further applications, such as semantic editing and expression transfer.

2. Related Work

Head correspondences. The concept of correspondence across head captures was introduced by early methods such as Blanz and Vetter [4] two decades ago, followed by extensive research on the optimization-based alignment of a common mesh topology to unstructured 3D head scans [2, 4, 24, 34, 46, 48, 54]. Recent learning-based approaches have moved towards predicting registered meshes directly from the raw scans [3, 40, 69], calibrated multi-view input images [5, 15, 31, 35, 41], or even monocular in-the-wild images [19]. While these works are restricted to geometry reconstructions, our method infers registered Gaussian splats representing both geometry and appearance.

Registered head scans enable the construction of 3D morphable models (3DMM) [4], which are used to establish temporal and spatial correspondences across single- or multi-view video frames through subject-specific head tracking [36, 52, 57]. Our method allows us to avoid this slow optimization process (12 s/frame for VHAP [52]) by directly predicting registered Gaussians in 0.5 s per frame.

Optimization-based head avatars. Current personalized head avatars require 3DMM tracking to establish correspondences before avatar optimization. These methods attach appearance representations, such as RGB textures [20], localized radiance field primitives [42], or Gaussian splats [18, 30, 64, 70], to the tracked geometry and optimize them against training images. GEM [70] shows that such high-quality avatars can be distilled into lightweight PCA-based representations. The high quality of the reconstructed avatars comes at the cost of long optimization times (45 h per avatar for GEM). We show that MATCH's predictions can be used to significantly accelerate the reconstruction of lightweight head avatars to 4.6 h per avatar.

Inference-based head avatars. One way to avoid the compute-heavy registration process is to infer animatable head avatars directly from one or a few images, typically at the cost of inferior synthesis quality. While early methods in this field focused on direct 2D image generation [6, 11, 55], recent works have moved towards 3D representations such as texturized meshes [26], Neural Radiance Fields [17, 50, 51], animatable triplanes [7, 10, 12, 13, 59], or 3DMM-attached Gaussian splats [21, 33, 63]. Similar to our approach, LAM [21] and FastAvatar [63] estimate Gaussians with a fixed UV location on a template mesh. However, their Gaussians are predicted in an unposed canonical space, while our method reconstructs the observed expression as-is in posed space with improved fidelity. Avat3r [29] and Facelift [44] predict pixel-aligned Gaussians [38, 56, 65, 67]. Unlike our method, these do not exhibit any cross-frame correspondences.

3. Method

This section presents MATCH, a feed-forward method for reconstructing photorealistic 3D heads as textures of registered Gaussian splats from calibrated multi-view images. We first describe the components of MATCH and then show how it can be applied to speed up the reconstruction process of lightweight, subject-specific head avatars (Section 3.2).

3.1. MATCH

Given V multi-view input images $\mathcal{I} \in \mathbb{R}^{H_{\text{img}} \times W_{\text{img}} \times 3}$ with known camera parameters, MATCH predicts a UV texture of Gaussian splat attributes $\mathcal{G}_{\{\mathbf{c},\boldsymbol{\alpha},\boldsymbol{\phi},\boldsymbol{\sigma},\boldsymbol{\theta}\}} \in \mathbb{R}^{H_{\text{uv}} \times W_{\text{uv}} \times C}$. Every texel is a C-vector that encodes RGB color $\mathbf{c}$, opacity $\boldsymbol{\alpha}$, location $\boldsymbol{\phi}$, anisotropic scale $\boldsymbol{\sigma}$, and rotation quaternion $\boldsymbol{\theta}$ of a Gaussian splat. An overview of MATCH's pipeline is presented in Figure 2.

Image tokenization. Following recent works [29, 67], we convert the high-resolution images $\mathcal{I}$ into image tokens $\mathcal{T}_{\text{img}}$ within low-resolution grids $\in \mathbb{R}^{H'_{\text{img}} \times W'_{\text{img}} \times d}$ that can be processed by the transformer. First, the images are concatenated with Plücker ray coordinates [49]. The result is split into patches and converted into tokens using a 2D convolutional layer with stride and kernel size equal to the patch size p_{img}. Following Avat3r [29], we fuse the image tokens with Sapiens [27] features through concatenation and linear projection, which yields the final tokens of dimension d.

Coarse mesh registration. Before computing the UV tokens, TEMPEH's [5] global stage without head localization is used to estimate a coarse mesh $\mathcal{M} \in \mathbb{R}^{N_{\text{vert}} \times 3}$ from the

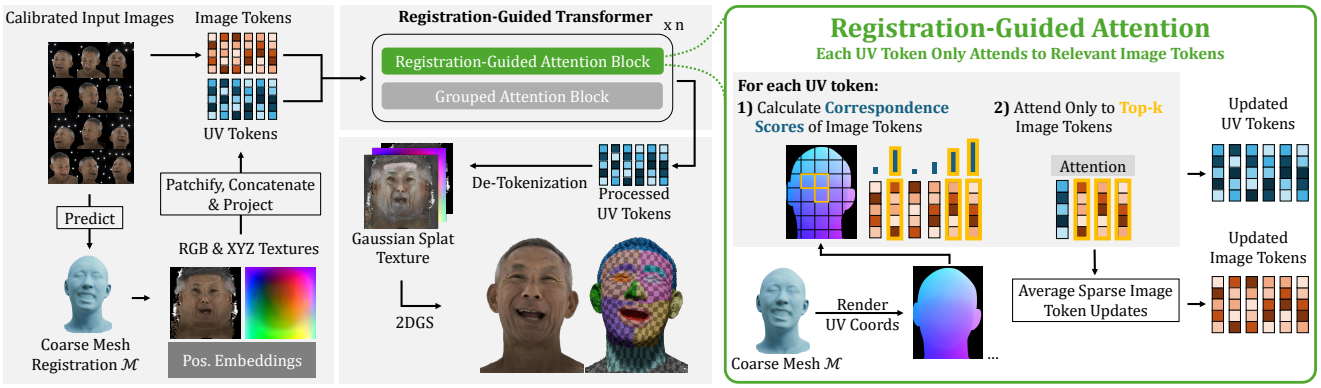


Figure 2. **Overview**. Given calibrated multi-view input images, MATCH first predicts a coarse mesh registration using a pretrained network. We obtain RGB and XYZ textures combined with learnable positional embeddings to encode UV tokens and follow GS-LRM [67] to tokenize the input images. The image and UV tokens serve as input to a transformer with two alternating attention blocks. In the novel registration-guided attention block, we render UV coordinate images from the input views, and for each UV token restrict the attention to image tokens displaying the relevant mesh region. The subsequent grouped attention block performs attention across the UV tokens and the tokens of each input image separately. The transformer outputs processed UV tokens that are projected into a texture of Gaussians.

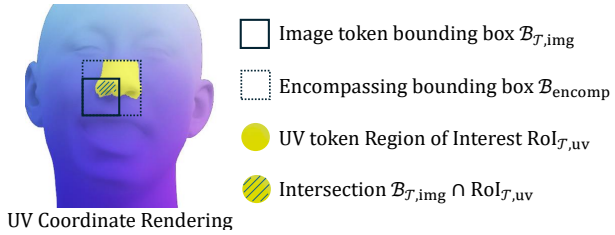


Figure 3. Correspondence score estimation between image tokens and UV tokens. To ease visualization, the full mesh is rasterized in overlay with the UV renders and patch sizes are increased.

input images. TEMPEH is trained against the ground-truth vertices of the Ava-256 dataset [45], including hair proxy geometry, and adopts the provided mesh topology.

UV tokenization. During UV tokenization, a high-resolution UV texture is encoded as a set of UV tokens $\mathcal{T}_{\text{uv}}$ in a low-resolution grid $\in \mathbb{R}^{H'_{\text{uv}} \times W'_{\text{uv}} \times d}$ that can be processed by the transformer. We first calculate a dense 3D location texture through per-texel barycentric interpolation of $\mathcal{M}$'s vertex locations. Second, we obtain an RGB texture by reprojecting the input images onto $\mathcal{M}$. These RGB and 3D location textures are concatenated and divided into non-overlapping patches of size p_{uv}. Finally, they are flattened, concatenated with learnable positional embeddings, and linearly projected to yield d-dimensional UV tokens.

Registration-guided attention. The image and UV tokens are processed by a transformer that contains alternating blocks of registration-guided attention and grouped attention. The registration-guided attention blocks constrain each UV token to only attend to image tokens of the corresponding head region (Figure 2 right). To determine which tokens should attend to each other, we estimate a correspondence score for each pair of UV and image tokens: We use the coarse mesh $\mathcal{M}$ to rasterize UV coordinates onto the input image planes. Since each UV token represents a texture patch, we can filter the rasterized UV coordinates by the respective coordinate range to obtain a region of interest $\text{RoI}_{\mathcal{T},\text{uv}}$ for that UV token, shown in yellow in Figure 3. Now let $\mathcal{B}_{\mathcal{T},\text{img}}$ denote the bounding box of a particular image token, and $\mathcal{B}_{\text{encomp}}$ the bounding box that includes both $\mathcal{B}_{\mathcal{T},\text{img}}$ and $\text{RoI}_{\mathcal{T},\text{uv}}$. The correspondence score S between the UV- and image token is defined as:

$$S\left(\mathcal{T}_{\text{uv}}, \mathcal{T}_{\text{img}}\right) = \frac{\text{RoI}_{\mathcal{T},\text{uv}} \cap \mathcal{B}_{\mathcal{T},\text{img}}}{\mathcal{B}_{\mathcal{T},\text{img}}} + \lambda \cdot \frac{\text{RoI}_{\mathcal{T},\text{uv}}}{\mathcal{B}_{\text{encomp}}}, \qquad (1)$$

where the arithmetic is performed on the respective areas in pixels. The first term measures the ratio of pixels within $\mathcal{B}_{\mathcal{T},\text{img}}$ coinciding with $\text{RoI}_{\mathcal{T},\text{uv}}$, while the second term accounts for image tokens in $\text{RoI}_{\mathcal{T},\text{uv}}$'s vicinity (weighted by $\lambda = 0.1$). After evaluating $S(\cdot)$ across all image tokens, for each UV token, we only attend to the $k_{\mathcal{T},\text{img}}$ highest-scoring image tokens. This localizes the reconstruction task, which improves generalization, and keeps the attention context length constant with an increasing number of input images, reducing computation costs for high image counts. After the attention operation, we obtain updated UV tokens, but only sparsely and potentially redundantly updated image tokens, which we address by averaging over all their occurrences.

Grouped attention. This block performs attention separately on the UV tokens and the image tokens of each input image, propagating information to unobserved head regions and enabling image-space feature processing with compute complexity scaling linearly in the number of input images.

UV de-tokenization. The transformer outputs processed UV tokens. We linearly project each token to a texture

patch of Gaussian parameters with shape $p_{\text{uv}} \times p_{\text{uv}} \times C$. These are assembled to the output Gaussian splat texture of shape $H_{\text{uv}} \times W_{\text{uv}} \times C$. For color and location, we follow Avat3r's [29] skip connections and apply the predictions as offsets to the initial values obtained during UV tokenization.

Training. Supervised by multi-view images and ground-truth meshes, MATCH is trained by minimizing:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{photometric}} + w_{\text{geometry}} \cdot \mathcal{L}_{\text{geometry}} + w_{\text{reg}} \cdot \mathcal{L}_{\text{reg}}. \quad (2)$$

The photometric loss $\mathcal{L}_{\text{photometric}}$ matches the predicted appearance against the ground-truth images,

$$\mathcal{L}_{\text{photometric}} = w_{\text{LPIPS}} \cdot \mathcal{L}_{\text{LPIPS}} + w_{\text{L1}} \cdot \mathcal{L}_{\text{L1}} + w_{\text{SSIM}} \cdot \mathcal{L}_{\text{SSIM}} \quad (3)$$

with perceptual loss [68] $\mathcal{L}_{\text{LPIPS}}$ and Structural Similarity Index Measure (SSIM) [62] $\mathcal{L}_{\text{SSIM}}$. We mask out the torso and upper shoulders using a pretrained semantic segmentation model [27]. $\mathcal{L}_{\text{geometry}}$ minimizes the L2-norm between the predicted Gaussian 3D locations and dense target locations from the ground-truth mesh registration. The regularization loss $\mathcal{L}_{\text{reg}}$ applies an L2 loss between the predictions and predefined target values for scale and opacity.

Implementation details: Our training consists of three stages. First, we train solely on the Ava-256 dataset [45], using $\mathcal{L}_{\text{geometry}}$ and $\mathcal{L}_{\text{reg}}$ only. In the second stage, we add $\mathcal{L}_{\text{photometric}}$. In the last stage, we train on a combination of the Ava-256 and the NeRSemble v2 dataset [28]. Since NeRSemble does not provide ground-truth geometry annotations, we deactivate $\mathcal{L}_{\text{geometry}}$ on samples drawn from it.

We adopt Ava-256's mesh topology and UV texture layout. MATCH predicts 64×64 UV tokens, each corresponding to 16×16 texture patches, totalling a 1024×1024 texture with 1M Gaussians that can be rendered at 570 fps [22]. The input images are divided into 8×8 patches.

Each training sample contains 12 head-centered images with a resolution of 640×512, which are used both as model input and rendering target. We train MATCH for 860k iterations on 4 NVIDIA H100 80GB GPUs, with per-GPU batch size 1, taking 11.8 days. Please see Section B for details.

3.2. Creation of Subject-Specific Head Avatars

Practical applications require animatable avatars. GEM [70] demonstrates the distillation of high-quality 3D head avatars (computationally and memory expensive) into a lightweight representation. We adapt this procedure to obtain controllable head avatars from sequences of Gaussian textures predicted by MATCH.

GEM. Given a multi-view video sequence, GEM performs mesh-based head tracking and optimizes a CNN-based head avatar [37], yielding a set of per-frame Gaussian parameter textures $\{\mathcal{G}_1, \ldots, \mathcal{G}_N\}$ in dense correspondence. Principal Component Analysis (PCA) is performed on these textures for each attribute—scale, position, opacity, and rotation—to build a linear Gaussian splat head model. Expressions are described via low-dimensional coefficients $\mathbf{k}_i$, which define linear combinations of the attribute-specific bases $\mathbf{B}_i$:

$$\mathcal{G} = \{\boldsymbol{\mu}_i + \mathbf{B}_i \mathbf{k}_i \mid i \in \{\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\theta}\}\}, \quad (4)$$

where the corresponding means are $\boldsymbol{\mu}_{\{\boldsymbol{\alpha},\boldsymbol{\phi},\boldsymbol{\sigma},\boldsymbol{\theta}\}}$. Colors are modeled as expression-independent constants. The bases are refined against the input images using a photometric loss. For image-based animation, two pretrained expression estimators [8, 14] extract features from input images, and a small MLP maps these features to the coefficients $\mathbf{k}_i$.

MATCH-based GEM avatars. The dense correspondence of the Gaussian splat textures predicted by MATCH allows us to bypass GEM's computationally expensive mesh tracking and CNN-based head avatar training. The predicted Gaussians are transformed into a canonical space through inverse linear blend skinning, followed by GEM's PCA decomposition, bases refinement, and MLP training steps. Contrasting GEM, we model dynamic color changes (except for the oral cavity), perform a joint PCA decomposition across all attributes, and optimize the PCA mean $\boldsymbol{\mu}_*$. Please refer to Section H for more details.

Dataset	Method	LPIPS ↓	CSIM ↑	PSNR ↑	SSIM ↑	L1 ↓	L2 ↓
Ava-256 [45]	GPAvatar [7]	0.303	0.507	21.304	0.714	0.046	0.009
	FastAvatar [63]	0.285	0.667	15.880	0.730	0.074	0.032
	LAM [21]	0.273	0.678	15.898	0.734	0.078	0.035
	Avat3r [29]	0.274	0.626	22.722	0.745	0.039	0.007
	FaceLift [44]	0.208	0.868	21.661	0.825	0.038	0.010
	Ours	0.163	0.928	23.680	0.848	0.027	0.008
NeRSemble [28]	GPAvatar [7]	0.259	0.599	25.296	0.801	0.029	0.003
	FastAvatar [63]	0.248	0.725	18.676	0.797	0.050	0.018
	LAM [21]	0.254	0.746	16.996	0.789	0.062	0.027
	FaceLift [44]	0.200	0.866	21.524	0.853	0.040	0.009
	Ours (Ava-256 only)	0.182	0.892	23.182	0.861	0.030	0.005
	Ours (NeRSemble only)	0.168	0.927	24.136	0.870	0.026	0.005
	Ours	0.152	0.944	25.509	0.884	0.024	0.003

Table 1. Novel view synthesis results on Ava-256 and NeRSemble.

4. Experiments

4.1. Datasets

We train and evaluate MATCH on Ava-256 [45] and NeRSemble v2 [28], which contain head-centric, multi-view video sequences of 256/425 subjects performing various facial expressions, captured by 80/16 cameras in a uniformly-lit studio environment. Ava-256 provides a 360° viewpoint coverage, whereas NeRSemble's cameras are restricted to $\pm 50°$ horizontally and $\pm 15°$ vertically. Ava-256 additionally includes registration meshes with hair proxy geometry for all frames. For validation, we held out 11/6 subjects on Ava-256 and NeRSemble.

4.2. Novel View Synthesis

We compare MATCH against single-view and multi-view 3D head reconstruction methods. LAM [21] and FastAvatar [63] employ transformers to predict 3DMM-attached

Figure 4. Novel view synthesis comparison on Ava-256 [45]. MATCH exhibits superior synthesis quality.

Gaussians while GPAvatar [7] generates animatable triplanes. Avat3r [29] and FaceLift [44] predict pixel-aligned Gaussian splats from the multi-view input images. We provide 12 input images at 640×512 resolution and render a random disjoint target view. For the single-image method LAM, the one with the smallest Euclidean distance to the target camera center is selected.

As shown in Figure 4, GPAvatar's reconstructions are blurry. FastAvatar and LAM achieve higher synthesis quality, but struggle with extreme expressions and exhibit identity shift. FaceLift and Avat3r produce more plausible results, yet FaceLift suffers from oversaturation artifacts and blurry textures, while Avat3r exhibits a loss of identity, especially in the mouth (top) and eye region (bottom). Overall, MATCH reconstructs the target subjects most faithfully, even for extreme expressions such as tongue protrusion. As CAP4D [58] only held out two Ava-256 subjects for testing, we only qualitatively compare to it on the single overlapping test subject in Figure 13. We find that MATCH reconstructs extreme expressions better than CAP4D.

Table 1 quantitatively compares MATCH on 1,000 samples from Ava-256 using standard metrics such as perceptual similarity (LPIPS), Peak Signal to Noise Ratio (PSNR), Structural Similarity Metric (SSIM), pixel-wise L1 and L2 errors, and the cosine similarity between identity vectors extracted by a face recognition network [9] (CSIM). We find that MATCH consistently outperforms all baselines.

Application to datasets w/o geometry. We train MATCH using the 3D head registrations available in Ava-256 as supervision. Although public datasets with larger subject counts exist [28, 47, 66], they lack registered meshes. This raises two questions: i) What is the impact of incorporating NeRSemble during training, despite the absence of ground-truth registrations? and ii) How well does MATCH generalize across datasets? To answer these questions, Figure 5 and Table 1 compare MATCH (trained jointly on Ava-256 and NeRSemble) against two model variants: one trained exclusively on Ava-256, and one trained exclusively on NeRSemble using pseudo-ground-truth meshes obtained via VHAP [52]. Note that the NeRSemble-only variant uses the FLAME mesh topology provided by VHAP, whereas the other versions utilize the Ava-256 topology. We evaluate these variants, alongside other baselines, on 1,000 NeRSemble samples. Even when trained solely on Ava-256, MATCH generates plausible synthesis results on the unseen NeRSemble dataset, outperforming all baselines. While training on NeRSemble alone improves performance, joint training on both datasets yields the best results, despite lacking geometry supervision for the NeRSemble samples. MATCH circumvents the need for expensive head registration on NeRSemble (estimated 216k GPU-hours for 65M frames) by enabling direct training on the images.

	LPIPS ↓	CSIM ↑	PSNR ↑	SSIM ↑	L1 ↓	L2 ↓
Dense Attention	0.221	0.849	20.364	0.794	0.041	0.013
w/o Sapiens	0.202	0.907	22.104	0.819	0.034	0.010
w/o Skipconn.	0.192	0.913	23.075	0.816	0.032	0.009
Orig. TEMPEH	0.190	0.909	22.775	0.823	0.033	0.009
$H_{\text{uv}} = W_{\text{uv}} = 256$	0.194	0.907	22.830	0.817	0.033	0.009
$H_{\text{uv}} = W_{\text{uv}} = 512$	0.190	0.914	22.829	0.822	0.032	0.009
Ours	0.187	0.918	23.032	0.825	0.032	0.009

Table 2. Ablation experiments on Ava-256.

Ablations. Table 2 evaluates the impact of individual model design choices and Figure 6 provides visualizations for the most influential components. All models are trained for 300k iterations on Ava-256. We find that the registration-guided attention yields the biggest performance gain; performing dense attention across all UV and image tokens ('Dense Attention') instead gives the worst scores. Qualitatively, we observe less fine detail in face accessories and worse hair textures with the dense attention approach. The pretrained Sapiens [27] feature extractor has the second-biggest impact. Disabling it ('w/o Sapiens') worsens generalization and reduces the synthesis quality.

Figure 5. Novel view synthesis on NeRSemble. Ours (Ava) / Ours (NeRSemble) are trained on Ava-256 and NeRSemble only, respectively.

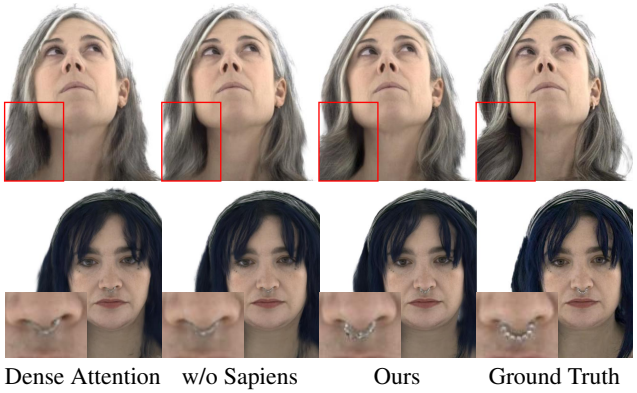


Figure 6. Ablation experiment on Ava-256.

Figure 7. Interpolation between Gaussian textures by factor γ.

Further, the reconstruction quality improves as we increase the UV texture resolution. Additionally, both removing the skip connections ('w/o Skipconn.') and using the original global TEMPEH model for geometry initialization instead of our adapted version ('Orig. TEMPEH'), produce slightly worse reconstructions. Section C.3 provides additional ablations on the number of input images and the number of image tokens $k_{\mathcal{T},\text{img}}$ to attend to in the registration-guided attention blocks. We find that MATCH produces plausible reconstructions from only four input images and that smaller $k_{\mathcal{T},\text{img}}$ improve performance.

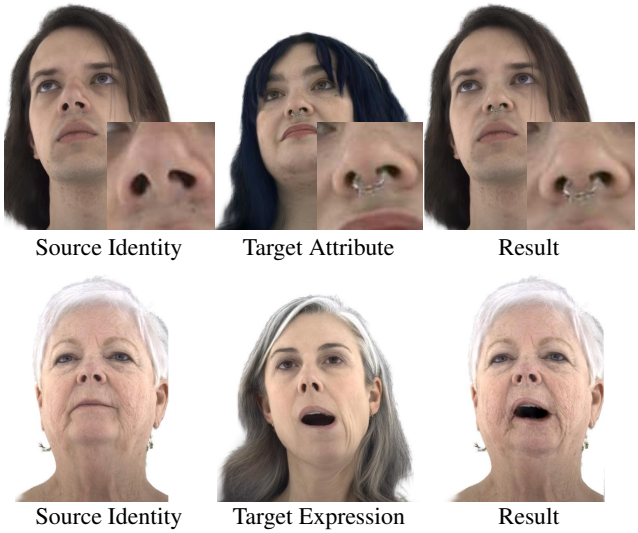


Figure 8. (Top) Semantic editing: Replacing a source's nose with that of a target. (Bottom) Expression transfer: Jaw articulation and mouth interior are transferred from a target to a source identity.

4.3. Interpolation, Editing & Expression Transfer

The dense semantic correspondence predicted by MATCH enables diverse applications, including cross-identity and cross-expression interpolation, part-based editing, and expression transfer. Figure 7 shows that simple interpolation between two subjects with different expressions gives smooth and plausible intermediate results. Figure 8 (top) demonstrates semantic editing, where Gaussian splat attributes in the nose region of a source identity are replaced with the ones from a target subject, resulting in a seamless blend. Additionally, Figure 8 (bottom) illustrates an arithmetic expression transfer approach, where the residual of Gaussian maps for an expressive and a neutral frame of a target subject is added to the neutral reconstruction of a source identity, resulting in a plausible expression transfer. Please refer to Figure 14 and Figure 20 for more results.

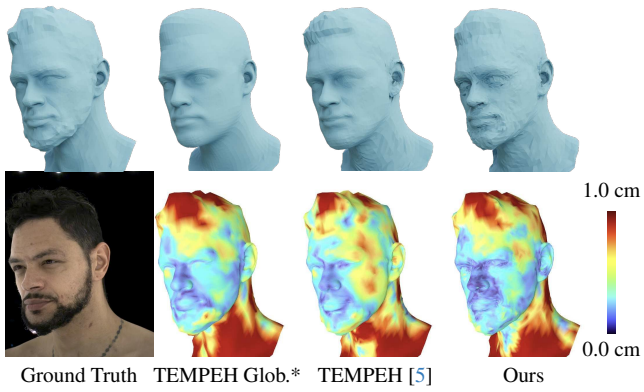


Figure 9. Geometry reconstruction comparison. The heatmaps visualize the Euclidean distance between the predicted and ground truth vertices. 'TEMPEH Glob.*' is the adapted version of TEMPEH's global stage that we use for the coarse mesh registration.

	Full		Face		Mouth		Eyes	
	P2P ↓	P2S ↓	P2P ↓	P2S ↓	P2P ↓	P2S ↓	P2P ↓	P2S ↓
TEMPEH Glob.*	7.34	2.76	3.84	1.50	2.69	1.15	2.27	1.09
TEMPEH [5]	7.84	2.18	4.04	0.97	2.99	0.69	2.66	0.62
Ours	6.69	2.10	3.18	0.88	2.14	0.72	1.54	0.61

Table 3. Quantitative geometry reconstruction comparison on Ava-256. Scores are reported in mm. P2P: Point-to-Point distance. P2S: Point-to-Surface distance. 'TEMPEH Glob.*' is the adapted version of TEMPEH's global stage that we use for the coarse mesh registration. 'Face' is the facial area w/o mouth, eyes, and ears.

4.4. Geometry Reconstruction

MATCH predicts texture maps of Gaussian splat attributes, including a map of 3D positions. Given vertex UV coordinates, these location textures can be converted into a mesh, making MATCH an inference-based head tracker. We compare our geometry reconstruction against TEMPEH [5], a state-of-the-art tracker that estimates registered meshes from multi-view images in a single inference step. For a fair comparison, we retrain TEMPEH on the Ava-256 dataset using the provided mesh topology. We also include the adapted TEMPEH global stage, which generates the coarse mesh registrations $\mathcal{M}$ used as input to our method. Table 3 reports the Euclidean vertex-to-vertex distance (P2P) and the Point-to-Surface (P2S) distance between the predicted and ground-truth meshes on 1,000 Ava-256 samples, while a qualitative comparison can be found in Figure 9. We observe that our method outperforms both baselines on both metrics when averaged over the entire head. In the mouth and eye regions, our method performs comparably to or slightly worse than TEMPEH. This is expected, as the ground truth meshes approximate the oral cavity and eyes with planar faces, while MATCH deviates from this coarse approximation to reconstruct the appearance necessary for a photorealistic rendering.

	Reconstr.	Self-Reenactment			Cross-Reenactment	
	Time ↓	LPIPS ↓	SSIM ↑	PSNR ↑	CSIM ↑	EmoL1 ↓
GA [53]	15.5h	0.233	0.755	21.248	0.629	10.618
RGBAvatar [32]	11.4h	0.233	0.782	22.185	0.657	10.333
GEM [70]	45.3h	0.214	0.778	21.761	0.800	9.866
Ours	4.6h	0.174	0.809	24.122	0.813	9.837

Table 4. Quantitative comparison of subject-specific head avatars.

4.5. Subject-Specific Head Avatars

We compare MATCH-based GEM avatars (Section 3.2) with GEM [70], GaussianAvatars (GA) [53], and RGBAvatar [32]. All methods are optimized per-subject on Ava-256 multi-view videos, captured by 12 cameras with an average sequence length of 3,300 frames. We evaluate the self- and cross-reenactment performances for five subjects on held-out sequences in Table 4 and Figure 10. For image-based animation, we drive GaussianAvatars and RGBAvatar with EMOCA's [8] parameters, which is also used as a pretrained feature extractor to drive GEM and our MATCH-based GEM avatar. Please refer to Section H.2 for more details. We measure standard image metrics for self-reenactment. For cross-reenactment, we evaluate the cosine similarity between features extracted with a pretrained face recognition network [9] (CSIM), and follow GEM [70] in evaluating the L1 distance between features of the emotion recognition network EmoNet [1] (EmoL1).

We find that our method consistently outperforms all baselines in the self-reenactment scenario and performs slightly better or on par with GEM for cross-reenactment. As we adapt GEM's image-based encoder, a similar cross-reenactment performance is plausible. RGBAvatar exhibits noticeable artifacts for subjects with long hair and produces blurry results for unseen expressions. GaussianAvatars generally handles unseen expressions more robustly, but is inferior to GEM and MATCH in terms of synthesis fidelity.

Our method not only improves synthesis quality but also drastically cuts reconstruction time. We achieve a $10\times$ reduction relative to GEM (the baseline with the closest synthesis quality) and are $2.5\times$ faster than the fastest baseline, RGBAvatar, which exhibits inferior visual fidelity. Please refer to Section H.3 and Section H.5 for detailed timings and ablation experiments on the changes compared to GEM.

5. Discussion

Registration dependency. MATCH uses Ava-256 mesh registrations as training supervision. While Section 4.2 shows cross-dataset training with mesh supervision on only one dataset, training entirely without it is subject to future work. This could potentially be achieved by training MATCH's geometry part solely on synthetic data.

Expression transfer. The direct transfer of extreme ex-

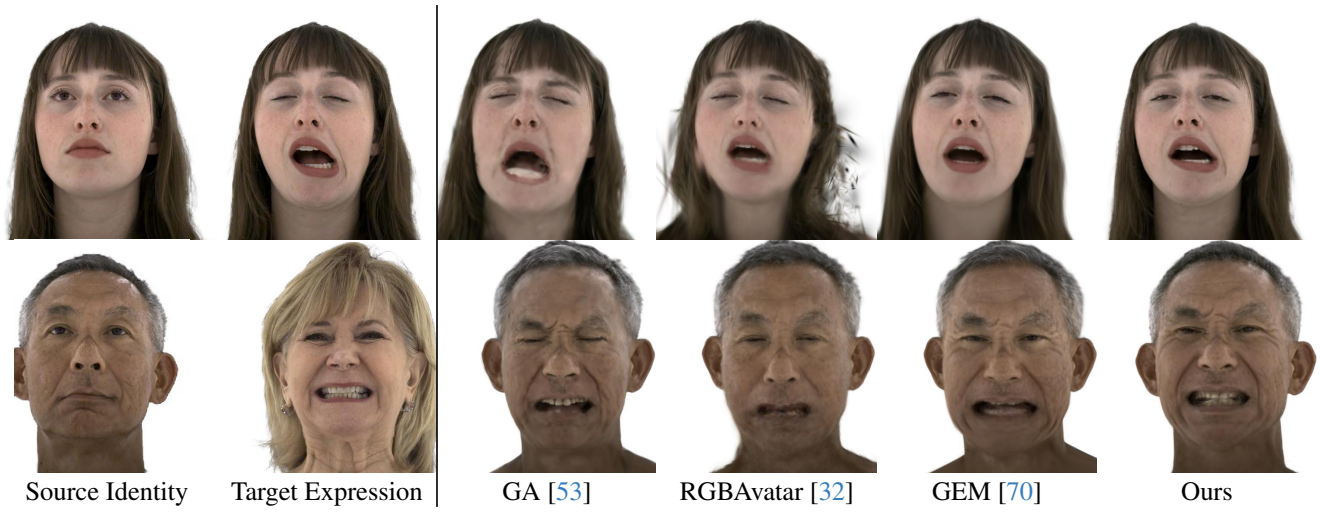


Figure 10. Qualitative comparison for image-based self-reenactment (top) and cross-reenactment (bottom) of the subject-specific avatars.

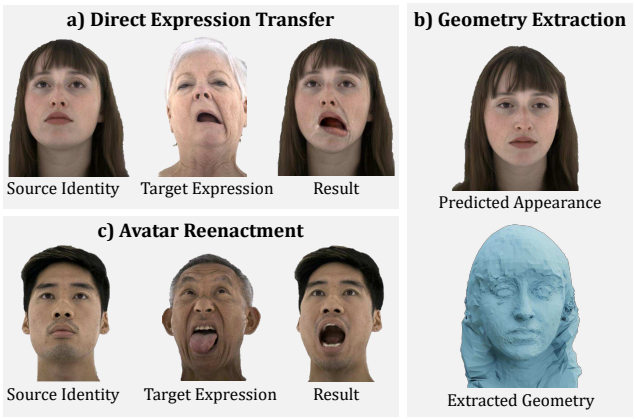


Figure 11. Limitations. a) Direct expression transfer (Section 4.3) may incur identity leakage for dissimilar identities. b) The meshes extracted from the predicted Gaussian 3D location textures (Section 4.4) can exhibit self-intersecting surfaces. c) The subject-specific head avatars (Section 3.2) are bound to interpolations of the training expressions and do not track eye movement.

pressions between MATCH's predicted Gaussian textures sometimes results in appearance leakage (Figure 11a). A less simplistic, learning-based method for expression transfer would likely be a more suitable choice.

Geometry reconstruction. Converting the predicted Gaussian 3D location textures into registered meshes occasionally results in self-intersecting surfaces (Section 4.4). While providing perfectly smooth registration meshes is not a primary goal of MATCH, adding additional surface regularization during training could mitigate such artifacts.

Avatar quality. By following GEM's [70] avatar formulation, the resulting avatars inherit its limitations of only generating expressions similar to those in the training data. Moving forward, an interesting direction is to leverage MATCH's cross-subject and cross-expression correspondence by learning a prior across identities and expressions, and then adapting it for per-subject avatars to achieve stronger generalization to unseen expressions.

6. Conclusion

We have presented MATCH, a framework that unifies geometry registration and radiance field reconstruction by predicting Gaussian splats in dense semantic correspondence. Central to this approach is our novel registration-guided attention mechanism, which enhances both computational efficiency and synthesis quality. By bypassing the bottleneck of traditional optimization-based tracking, MATCH reduces the total avatar creation time from 45 hours to just 4.6 hours, a $10\times$ speedup over the qualitatively closest baseline, while simultaneously improving visual fidelity. Beyond efficiency, the learned intra- and inter-subject correspondence unlocks diverse applications, including semantic editing, robust geometry registration, and identity interpolation. Overall, MATCH represents a clear step toward scalable, high-fidelity avatar creation and the structured Gaussian registration representation provides a solid basis for future work on learning priors across different subjects.

7. Acknowledgements

We thank W. Zielonka for data preparation and GEM training support, and B. B. Bilecen, B. Thambiraja, A.-L. Schweikert, S. Kocour, P.-W. Grassal, X. Lyu, J. Thies, and F. Rajič for proofreading. MP was partially funded by the Max Planck ETH Center for Learning Systems (CLS). Computations were performed on the MPI-IS Tübingen cluster. The MATCH icon is from flaticon.

References

[1] Muhammad Abdul-Mageed and Lyle Ungar. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 718–728, 2017. 7

[2] Brian Amberg, Reinhard Knothe, and Thomas Vetter. Expression invariant 3d face recognition with a morphable model. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE, 2008. 2

[3] Mehdi Bahri, Eimear O'Sullivan, Shunwang Gong, Feng Liu, Xiaoming Liu, Michael M Bronstein, and Stefanos Zafeiriou. Shape my face: registering 3d face scans by surface-to-surface translation. *International Journal of Computer Vision*, 129(9):2680–2713, 2021. 2

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999. 2

[5] Timo Bolkart, Tianye Li, and Michael J. Black. Instant multi-view head capture through learnable registration. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 768–779, 2023. 2, 7, 6

[6] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020. 2

[7] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. GPAvatar: Generalizable and precise head avatar from image(s). In *The Twelfth International Conference on Learning Representations*, 2024. 2, 4, 5, 6, 9, 10

[8] Radek Daněček, Michael J Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 4, 7

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 5, 7

[10] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7119–7130, 2024. 2

[11] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International conference on Computer Vision*, pages 14398–14407, 2021. 2

[12] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022. 2

[13] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8498–8507, 2024. 2

[14] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *Transactions on Graphics, (Proc. SIGGRAPH)*, 40(8), 2021. 4

[15] Panagiotis Filntisis, George Retsinas, Radek Danecek, Vanessa Sklyarova, Petros Maragos, and Timo Bolkart. MOCHI: Registration-free learnable multi-view capture of faces in dense semantic correspondence. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 2

[16] Alex Fisher, Ricardo Cannizzaro, Madeleine Cochrane, Chatura Nagahawatte, and Jennifer L Palmer. COLMAP: A memory-efficient occupancy grid mapping framework. *Robotics and Autonomous Systems*, 142:103755, 2021. 5

[17] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2021. 2

[18] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Npga: Neural parametric gaussian avatars. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1, 2

[19] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction. *arXiv preprint arXiv:2505.00615*, 2025. 2

[20] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18653–18664, 2022. 2

[21] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–13, 2025. 2, 4, 5, 6, 9, 10

[22] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D Gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024. 1, 4

[23] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992. 6

[24] Tim J Hutton, BR Buxton, and Peter Hammond. Dense surface point distribution models of the human face. In *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2001)*, pages 153–160. IEEE, 2001. 2

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time

radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[26] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *European Conference on Computer Vision*, pages 345–362. Springer, 2022. 2

[27] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024. 2, 4, 5

[28] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 4, 5, 1, 2, 10

[29] Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12089–12100, 2025. 2, 4, 5, 1, 9

[30] Junxuan Li, Chen Cao, Gabriel Schwartz, Rawal Khirodkar, Christian Richardt, Tomas Simon, Yaser Sheikh, and Shunsuke Saito. Uravatar: Universal relightable gaussian codec avatars. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2

[31] Jing Li, Di Kang, and Zhenyu He. GRAPE: generalizable and robust multi-view facial capture. In *European Conference on Computer Vision (ECCV)*, pages 403–418. Springer, 2024. 2

[32] Linzhou Li, Yumeng Li, Yanlin Weng, Youyi Zheng, and Kun Zhou. Rgbavatar: Reduced gaussian blendshapes for online modeling of head avatars. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10747–10757, 2025. 7, 8, 12, 13

[33] Peng Li, Yisheng He, Yingdong Hu, Yuan Dong, Weihao Yuan, Yuan Liu, Siyu Zhu, Gang Cheng, Zilong Dong, and Yike Guo. Panolam: Large avatar model for gaussian full-head synthesis from one-shot unposed image. *arXiv preprint arXiv:2509.07552*, 2025. 2

[34] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2, 6

[35] Tianye Li, Shichen Liu, Timo Bolkart, Jiayi Liu, Hao Li, and Yajie Zhao. Topologically consistent multi-view face inference using volumetric sampling. In *International Conference on Computer Vision (ICCV)*, pages 3824–3834, 2021. 2

[36] Xuanchen Li, Yuhao Cheng, Xingyu Ren, Haozhe Jia, Di Xu, Wenhan Zhu, and Yichao Yan. Topo4d: Topology-preserving gaussian splatting for high-fidelity 4d head capture. In *European Conference on Computer Vision*, pages 128–145. Springer, 2024. 2

[37] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4

[38] Chenguo Lin, Panwang Pan, Bangbang Yang, Zeming Li, and Yadong Mu. Diffsplat: Repurposing image diffusion models for scalable 3d gaussian splat generation. In *International Conference on Learning Representations (ICLR)*, 2025. 2

[39] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. *arXiv*, pages arXiv–2012, 2020. 1

[40] Feng Liu, Luan Tran, and Xiaoming Liu. 3d face modeling from diverse raw scan data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9408–9418, 2019. 2

[41] Shichen Liu, Yunxuan Cai, Haiwei Chen, Yichao Zhou, and Yajie Zhao. Rapid face asset acquisition with recurrent feature alignment. *Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 41(6):214:1–214:17, 2022. 2

[42] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 2

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[44] Weijie Lyu, Yi Zhou, Ming-Hsuan Yang, and Zhixin Shu. Facelift: Learning generalizable single image 3d face reconstruction from synthetic heads. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12691–12701, 2025. 2, 4, 5, 6, 9, 10

[45] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024. 3, 4, 5, 1, 2, 6, 9, 10

[46] Iordanis Mpiperis, Sotiris Malassiotis, and Michael G Strintzis. Bilinear models for 3-d face and facial expression recognition. *IEEE Transactions on Information Forensics and Security*, 3(3):498–511, 2008. 2

[47] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin WANG, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, and Kwan-Yee Lin. Renderme-360: A large

digital asset library and benchmarks towards high-fidelity head avatars. In *Advances in Neural Information Processing Systems*, pages 7993–8005. Curran Associates, Inc., 2023. 5

[48] Georgios Passalis, Panagiotis Perakis, Theoharis Theoharis, and Ioannis A Kakadiaris. Using facial symmetry to handle pose variations in real-world 3d face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1938–1951, 2011. 2

[49] Julius Plucker. Xvii. on a new geometry of space. *Philosophical Transactions of the Royal Society of London*, (155): 725–791, 1865. 2

[50] Malte Prinzler, Otmar Hilliges, and Justus Thies. DINER: Depth-aware image-based neural radiance fields. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[51] Malte Prinzler, Egor Zakharov, Vanessa Sklyarova, Berna Kabadayi, and Justus Thies. Joker: Conditional 3d head synthesis with extreme facial expressions. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2025. 2

[52] Shenhan Qian. Vhap: Versatile head alignment with adaptive appearance priors, 2024. 2, 5, 1, 7

[53] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic head avatars with rigged 3D gaussians. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20299–20309, 2024. 1, 7, 8, 12, 13

[54] Augusto Salazar, Stefanie Wuhrer, Chang Shu, and Flavio Prieto. Fully automatic expression-invariant face correspondence. *Machine Vision and Applications*, 25(4):859–879, 2014. 2

[55] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 2

[56] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 2

[57] Felix Taubner, Prashant Raina, Mathieu Tuli, Eu Wern Teh, Chul Lee, and Jinmiao Huang. 3D face tracking from 2D video through iterative dense UV to image flow. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1227–1237, 2024. 2

[58] Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B. Lindell. CAP4D: Creating animatable 4D portrait avatars with morphable multi-view diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5318–5330, 2025. 5, 2, 3, 8

[59] Phong Tran, Egor Zakharov, Long-Nhat Ho, Liwen Hu, Adilbek Karmanov, Aviral Agarwal, McLean Goldwhite, Ariana Bermudez Venegas, Anh Tuan Tran, and Hao Li. Voodoo xp: Expressive one-shot head reenactment for vr telepresence. *ACM Transactions on Graphics, Proceedings of the 17th ACM SIGGRAPH Conference and Exhibition in Asia 2024, (SIGGRAPH Asia 2024), 12/2024*, 2024. 2

[60] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: visual geometry grounded transformer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306, 2025. 2

[61] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. 2

[62] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[63] Yue Wu, Yufan Wu, Wen Li, Yuxi Lu, Kairui Feng, and Xuanhong Chen. Fastavatar: Towards unified fast high-fidelity 3d avatar reconstruction with large gaussian reconstruction transformers. *arXiv preprint arXiv:2508.19754*, 2025. 2, 4, 5, 6, 9, 10

[64] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[65] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision*, pages 1–20. Springer, 2024. 2

[66] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[67] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3D Gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 2, 3

[68] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 4

[69] Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. Imface: A nonlinear 3d morphable face model with implicit neural representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20343–20352, 2022. 2

[70] Wojciech Zielonka, Timo Bolkart, Thabo Beeler, and Justus Thies. Gaussian eigen models for human heads. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15930–15940, 2025. 1, 2, 4, 7, 8, 6, 12, 13

Feed-forward Gaussian Registration for Head Avatar Creation and Editing

Supplementary Material

Figure 12. MATCH predictions overlaid with texturized Gaussian splat renderings.

Contents. This supplementary material provides additional results, implementation details, and analyses that support the claims presented in the main paper. Section A gives more information on the datasets used for training and evaluation, followed by implementation details in Section B. Section C provides more novel view synthesis experiments. Section D and Section E present reconstructions for in-the-wild and single-input-image scenarios respectively. Section F shows additional results for the interpolation, semantic editing, and expression transfer on Gaussian splat textures. Section G supplements a quantitative evaluation of the correspondences predicted by MATCH. Section H adds further comparisons of the subject-specific head avatars created from MATCH's predictions. We close with ethical considerations in Section I.

A. Datasets

Ava-256. The Ava-256 dataset [45] provides multi-view video captures from 80 cameras of 256 subjects performing a wide range of expressions. Ava-256 comes with ground truth mesh registrations of the head, including hair proxy geometry. We sample the captured videos at a framerate of 0.75 fps and select cameras that are evenly distributed over the frontal hemisphere in a range of $\pm 40°$ horizontally and $[-15°, +36°]$ vertically. To facilitate generalization to other datasets with different backgrounds, we remove the image background using the provided alpha masks. We adopt Avat3r's [29] train-validation split and use 244 identities for training and 11 for validation. One subject was removed from Avat3r's validation set due to faulty ground truth segmentation masks. We sample 123,000 frames for training and 1,000 for validation.

NeRSemble. The NeRSemble v2 dataset [28] provides multi-view video captures from 16 cameras of 425 subjects, which are split into 419 training and 6 validation subjects. We select cameras

```
221501007, 222200045,
222200049, 222200043,
222200047, 222200038,
220700191, 222200041,
222200046, 222200040,
222200042, 222200044
```

for training and leave the rest for validation. As with Ava-256, we sample 123,000 frames for training and 1,000 for validation. Further, we use a pretrained matting model [39] to whiten the background. Pseudo ground truth mesh registrations are obtained with an optimization-based head tracker [52].

B. Implementation Details

Table 6 provides a list of the most important hyperparameter values, and Table 5 presents the loss weights used in the different training stages. We use the AdamW optimizer [43] with an initial learning rate of $4e-5$, a weight decay of 0.05, and a cosine learning rate scheduler that decreases the

Iteration	Stage Name	Training Datasets	Loss Weights
0 - 100k	Geometry Only	Ava-256 [45]	$w_{\text{geometry}} = 1 \times 10^{-3}$ $w_{\text{reg}} = 1 \times 10^{-3}$ $w_{\text{lpips}} = 0$ $w_{\text{L1}} = 0$ $w_{\text{SSIM}} = 0$
100k - 400k	Geometry & Apparance	Ava-256 [45]	$w_{\text{geometry}} = 1 \times 10^{-3}$ $w_{\text{reg}} = 1 \times 10^{-3}$ $w_{\text{L1}} = 0.8$ $w_{\text{SSIM}} = 0.2$ $w_{\text{LPIPS}} = 0$ until 150k, then linearly increasing to 1.0 until 200k
400k - 860k	Mixed Training	Ava-256 [45] & NeRSemble [28]	$w_{\text{geometry}} = (\text{Ava-256: } 1 \times 10^{-3}; \text{NeRSemble: } 0)$ $w_{\text{reg}} = 1 \times 10^{-3}$ $w_{\text{L1}} = 0.8$ $w_{\text{SSIM}} = 0.2$ $w_{\text{LPIPS}} = 1$

Table 5. MATCH loss weights in different training stages.

Parameter	Value
UV texture resolution $H_{\text{uv}} \times W_{\text{uv}}$	1024×1024
UV token patch size p_{uv}	16
Image token patch size p_{img}	8
Token dimension d	512
Number of input images V	12
Image resolution $H_{\text{img}} \times W_{\text{img}}$	640×512
Number of registration-guided attention blocks	6
Registration-guided attention image token count $k_{\mathcal{T},\text{img}}$	100
Number of grouped attention blocks	6
Gaussian scale regularization target	5×10^{-4}
Gaussian opacity regularization target	0.7
Learnable UV token positional embedding dimension	512

Table 6. MATCH hyperparameters.

learning rate to 0 within 1M steps after a 1,000-step linear warm-up phase. For calculating the Sapiens feature maps of the input images, we assemble them into grids of 2×2 before feeding them into the feature extractor to save computation time.

C. Novel View Synthesis

C.1. Detailed Baseline Description

We compare our model against the baselines GPAvatar [7], Fastavatar [63], LAM [21], Avat3r [29], FaceLift [44], and CAP4D [58].

LAM [21] predicts Gaussian splats for each vertex of a subsampled 3DMM from a single input image using a transformer backbone. These can be directly driven using 3DMM parameters. Fastavatar [63] builds on this approach and enables the aggregation of information extracted from several input images. GPAvatar [7] follows a different approach and reconstructs 3D head avatars from one or several input images using a triplane representation that can be animated with point-based expression fields.

FaceLift [44] trains a Gaussian Splatting Large Reconstruction Model (GS-LRM) [67] on synthetic data to predict pixel-aligned Gaussian splats from several input images. This GS-LRM is used to lift predictions of a diffusion model, which infers multi-view images from a single reference, into a 3D Gaussian splatting representation. In our comparison, we solely focus on the GS-LRM model of FaceLift, which receives the ground truth multi-view images as input. Avat3r [29] similarly regresses pixel-aligned Gaussian splats, yet they can be directly animated into new expressions through cross-attention to latent expression codes. Note that these latent expression codes are constructed from high-quality mesh registrations and texture reprojections that are obtained with closed-source software, making Avat3r inapplicable to datasets other than Ava-256. The Gaussian splats predicted by FaceLift and Avat3r are pixel-aligned and do not exhibit any semantic correspondence across frames or subjects.

CAP4D [58] uses a 3DMM-conditioned multi-view diffusion model to generate images with novel pose and expression, given one or several input images. In contrast to the other baselines, which infer 3D representations, it predicts 2D images that are not truly 3D-consistent. CAP4D only held out two subjects of the Ava-256 dataset for validation, only one of which intersects with the validation subjects from Avat3r and our method. As a consequence, we only perform a qualitative comparison with CAP4D on this one subject (see Figure 13).

Since different methods predict different crops of the face, we evaluate the methods on the maximum square crop that fits into the intersection of all bounding boxes, resized to a resolution of 512×512, and mask out the torso and shoulders using a pretrained segmentation network [27].

CAP4D [58] Ours Ground Truth

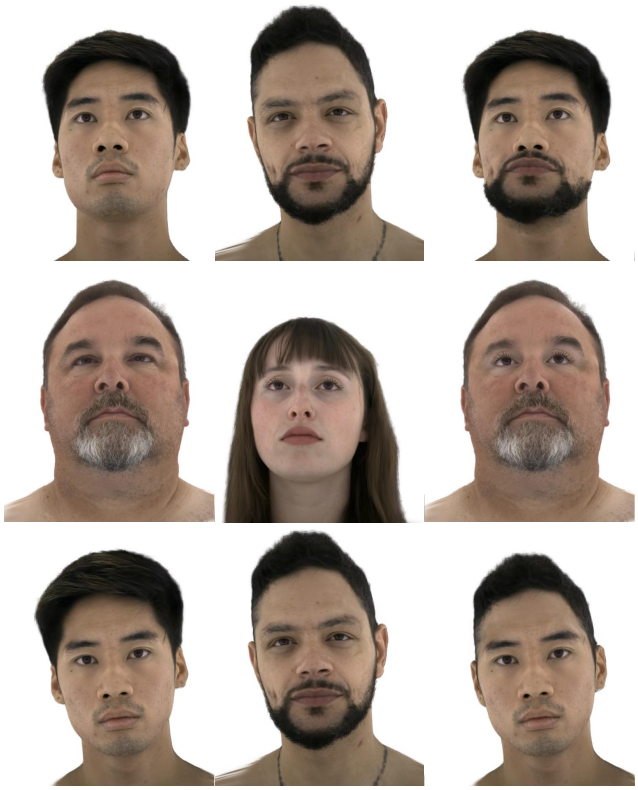
Figure 13. Novel view synthesis comparison against CAP4D on Ava-256.

V_{TEMPEH}	V_{MATCH}	LPIPS ↓	CSIM ↑	PSNR ↑	SSIM ↑	L1 ↓	L2 ↓
12	2	0.213	0.866	20.503	0.795	0.042	0.013
12	4	0.212	0.865	21.078	0.798	0.039	0.012
12	8	0.195	0.907	22.350	0.816	0.034	0.010
12	12	0.187	0.918	23.032	0.825	0.032	0.009
12	16	0.182	0.923	23.367	0.831	0.032	0.009
2	2	0.261	0.689	15.996	0.731	0.071	0.033
4	4	0.230	0.798	19.289	0.775	0.047	0.017
8	8	0.198	0.903	22.076	0.812	0.035	0.010
12	12	0.187	0.918	23.032	0.825	0.032	0.009
16	16	0.182	0.924	23.265	0.831	0.032	0.009

Table 7. Quantitative ablation of the number of input views to MATCH evaluated on Ava-256. We evaluate two scenarios: i) Changing the number of input views to MATCH while keeping the number of inputs to the coarse mesh registration model (TEMPEH) at the default ($V = 12$). ii) Changing the number of input views for both TEMPEH and MATCH.

C.2. Further qualitative comparisons

Figure 13 presents the qualitative comparison with CAP4D on the one intersecting validation subject. We observe that our method predicts reconstructions with better identity preservation and expression fidelity. Figure 23 and Figure 24 provide additional comparisons with the remaining baselines on samples from Ava-256 and NeRSemble respectively. As discussed in the main paper, our method exhibits superior synthesis quality.



Source Identity Target Attribute Result

Figure 14. Additional semantic editing results. From top to bottom: Transferring beard and lips, eyes, and hairstyle.

C.3. Additional Ablations

Number of input images. Figure 15 qualitatively evaluates the impact of the number of input views on the synthesis result. We conduct two lines of experiments: i) Keeping the number of input images to the coarse mesh registration model (TEMPEH) at the default ($V = 12$) while only changing the number of input views to MATCH. This evaluates the actual impact of the number of input views on our method in isolation. ii) TEMPEH and MATCH receive the same number of input images. This is the more realistic scenario, but entangles the sensitivity of TEMPEH to few input images with MATCH's.

We find that MATCH is highly robust to few input images and can generate plausible reconstructions even for two input images, assuming high-quality geometry initialization. However, TEMPEH's geometry prediction degrades significantly for two views, resulting in a low-quality reconstruction for the combined scenario. If TEMPEH and MATCH receive the same number of input views, starting from four images, plausible results are produced. Fine details improve as more input views are added. This is confirmed by Table 7 and Figure 17 (top), which show improving LPIPS scores as the number of views increases.

Figure 15. Qualitative ablation study for the number of input views to MATCH. We evaluate two scenarios. Top: Changing the number of input views to MATCH while keeping the number of inputs to the coarse mesh registration model (TEMPEH) at the default ($V = 12$). Bottom: Changing the number of input views for both TEMPEH and MATCH.

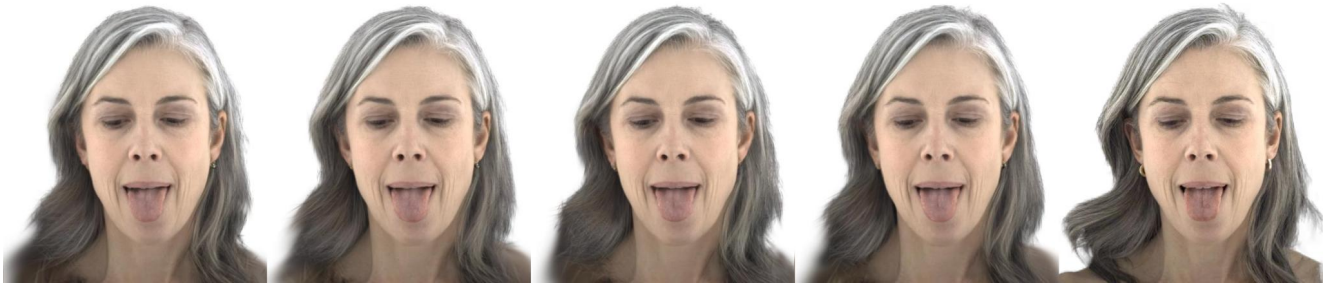


Figure 16. Qualitative ablation study for $k_{\mathcal{T},\text{img}}$, i.e., the number of image tokens that each UV token attends to in the registration-guided attention blocks. The default value is $k_{\mathcal{T},\text{img}} = 100$.

Figure 15 (bottom) reports the inference speed. As discussed in the main paper, while the computational complexity scales quadratically with the number of input images for dense attention between all UV and image tokens, our method's complexity increases only linearly. Especially for high numbers of input images, this results in a considerable improvement of inference speed ($1.8\times$ acceleration compared to dense attention for 16 input images). We found 12 input images to be a good compromise between inference speed and synthesis quality, running at a framerate of 2 fps.

Registration-guided attention context length. We ablate the effect of $k_{\mathcal{T},\text{img}}$, i.e., the number of image tokens that each UV token attends to in the registration-guided attention blocks. The quantitative comparison in Table 8 shows minor improvements as we decrease $k_{\mathcal{T},\text{img}}$. However, we did not observe pronounced qualitative differences as shown in Figure 16.

Robustness to coarse mesh errors. MATCH uses a coarse mesh estimated by TEMPEH as initialization. In practice, TEMPEH exhibits moderate inaccuracies, which MATCH can recover from, see Figure 18 (left), producing plausible results. When perturbing the coarse mesh by constant vertex offsets Δx, Figure 18 (right), artifacts only appear for $\Delta x \geq 20$mm, which is $7\times$ higher than the average point-to-surface distance of TEMPEH.

Image token self attention. MATCH performs self-

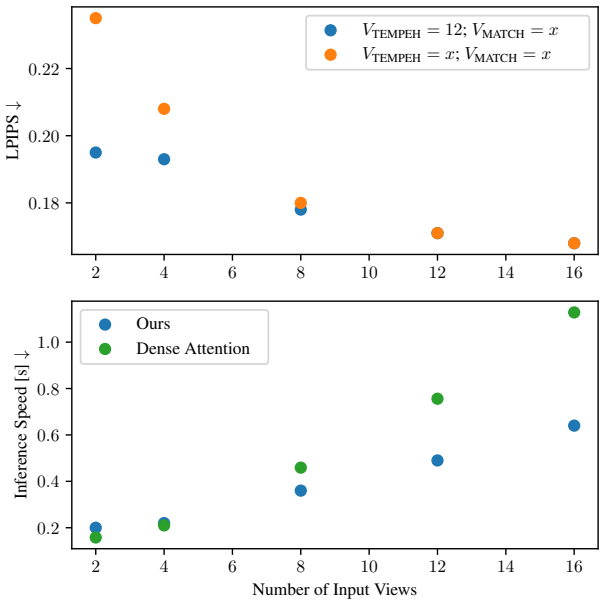
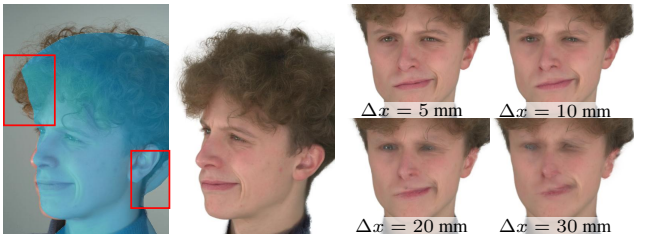


Figure 17. Top: Quantitative ablation study for the number of input views to MATCH on Ava-256. We evaluate two scenarios: i) Changing the number of input views to MATCH while keeping the number of inputs to the coarse mesh registration model (TEMPEH) at the default ($V = 12$). ii) Changing the number of input views for both TEMPEH and MATCH. Bottom: Inference speed comparison between our model with the novel registration-guided attention versus a version with dense attention across all UV and image tokens.



GT Image & Coarse Mesh Prediction Predictions For Perturbed Coarse Meshes

Figure 18. Robustness to errors in the coarse TEMPEH mesh.

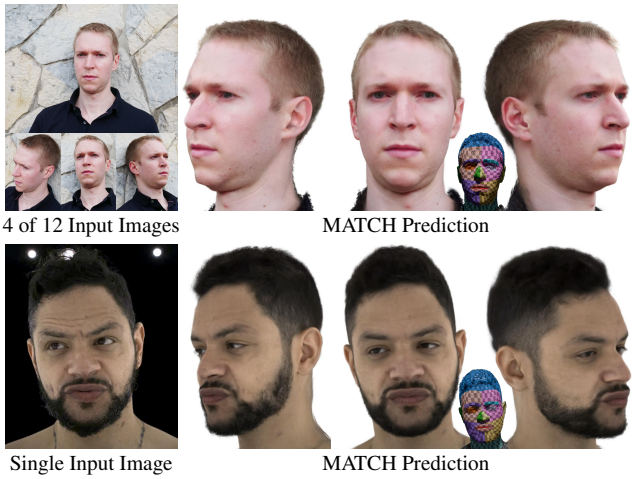
attention between the image tokens of each image to enable image-space signal processing inside the grouped attention blocks. We found that this self-attention can be skipped without harming the model performance while increasing the inference speed by 8%.

D. In-the-wild application.

While MATCH was trained on calibrated studio-captures with uniform lighting and known camera parameters, we found that it generalizes to in-the-wild captures and yields high-quality reconstructions, see Figure 19 (top). The input images were captured with an off-the-shelf camera in an outdoor environment, and we used COLMAP [16] to estimate the camera parameters.

$k_{\mathcal{T},\text{img}}$	LPIPS ↓	CSIM ↑	PSNR ↑	SSIM ↑	L1 ↓	L2 ↓
25	0.184	0.926	22.908	0.830	0.032	0.009
50	0.183	0.924	23.164	0.830	0.031	0.008
100	0.187	0.918	23.032	0.825	0.032	0.009
150	0.185	0.919	22.951	0.826	0.032	0.009

Table 8. Quantitative ablation study for $k_{\mathcal{T},\text{img}}$, i.e., the number of image tokens that each UV token attends to in the registration-guided attention blocks. The evaluations were performed on the Ava-256 dataset. The default value is $k_{\mathcal{T},\text{img}} = 100$.



4 of 12 Input Images MATCH Prediction

Single Input Image MATCH Prediction

Figure 19. In-the-wild (top) and single-image (bottom) results.

E. Single-image inference.

MATCH is not trained to hallucinate unobserved regions and shows artifacts for two input images only (see Figure 15). However, we can follow FaceLift [44] to generate additional views from a single input image with the 2D prior of CAP4D and input these to MATCH. Figure 19 (bottom) shows that this yields high-quality reconstructions.

F. Additional Results for Interpolation, Editing, and Expression Transfer

Figure 25, Figure 14, Figure 20 present further results for interpolation, semantic editing, and expression transfer, respectively. We observe smooth interpolations between samples and plausible editing results for swapping beard, eyes, and hairstyle. As discussed in the main paper, the arithmetic expression transfer approach, where the residual of Gaussian maps for an expressive and a neutral frame of a target subject is added to the neutral reconstruction of a source identity, can result in uncanny results for extreme expressions and dissimilar identities. A less simplistic method,

Corresp. Dist. [mm] ↓	Full Head	Face	Ears	Eyes	Mouth	Scalp
TEMPEH [5] / Ours	8.9 / **8.0**	5.4 / **4.8**	10.8 / **9.1**	2.5 / **2.1**	2.8 / **2.7**	13.5 / **12.5**

Table 9. Quantitative correspondence evaluation.

Figure 20. Additional expression transfer results. Note that we only aim to transfer the oral expression and do not apply any modifications to other regions, e.g., the eyes.

e.g., a conditional VAE [45], would be a more suitable choice for this challenging task.

G. Quantitative Correspondence Evaluation.

We quantify the semantic correspondence of MATCH 's predictions using Ava-256's ground truth mesh registrations. Table 9 reports the Euclidean distance between the center of each predicted Gaussian and its corresponding target location obtained through barycentric interpolation on 1,000 samples. The same interpolation is done to evaluate TEMPEH's results. We find that MATCH produces superior correspondence.

H. Additional Material for the Subject-Specific Avatars

H.1. Detailed Avatar Creation Procedure

This section illustrates the changes applied to GEM's [70] procedure to create a lightweight animatable avatar from a set of Gaussian splat textures predicted by MATCH. Ablations on the effect of the individual changes are presented in Figure 22 and Table 11, which are discussed in Section H.5. Figure 21 provides an overview of the resulting procedure.

i) Skip Tracking & CNN-based Avatar Training: Since MATCH directly predicts Gaussian splats that are in correspondence across frames, we can skip the time-expensive procedure of tracking and CNN-based head avatar training, which drastically reduces the time to create a lightweight head avatar (see Table 10). Since the reconstruction of the PCA basis requires unposed Gaussians in a canonical space, we have to unpose MATCH's predictions. To this end, we extract the Ava-256 mesh by sampling the texture of predicted 3D Gaussian locations at the template vertices' UV coordinates. We then convert the Ava-256 mesh to the topology of FLAME [34], a publicly available 3D morphable model (3DMM), using a fixed mapping of vertex locations. The 3DMM parameters are obtained by optimizing FLAME's vertices against our vertex predictions using a Huber loss [23]. Finally, we can use the obtained FLAME pose parameters to apply inverse linear blend skinning to transform the Gaussians predicted by MATCH into an unposed canonical space on which we perform the PCA decomposition. Note that during this unposing operation, the jaw articulation is neutralized as well. For this reason, we train GEM's expression encoder to also predict the jaw pose in addition to the Eigen-coefficients. To reduce the compute cost and memory requirements of the PCA decomposition, we use a version of MATCH that predicts Gaussian textures with a reduced resolution of 512×512.

ii) Modality-agnostic PCA: GEM creates separate PCAs for each of the Gaussian's modalities (rotation, position, opacity, and scale). However, we found that this formulation misses crucial correlations between the modalities (e.g., raised eyebrows should correlate with color changes in wrinkles on the forehead). This is resolved by modelling all Gaussian modalities in a joint PCA.

iii) Enable dynamic colors: GEM disables dynamic color changes to promote semantic correspondence of Gaussians across frames. For MATCH, however, this is neither feasible nor practical, since it must predict dynamically changing colors to reconstruct the appearance of different subjects, and intrinsically exhibits high semantic correspondence across subjects and frames. As such, we drop the constraint of static colors during the PCA reconstruction and refinement. The only exception to this is the interior of the mouth cavity. Since the ground truth mesh registrations on the Ava-256 dataset simplify the oral cavity as planar surfaces between the lips, the semantic correspondence of MATCH's predictions in this area is limited. Plausible reconstructions are achieved through intricately changing colors, opacities, and scales. We found that naïvely using the lightweight expression MLP to predict these complex dynamics yields test-time artifacts. To alleviate this problem,

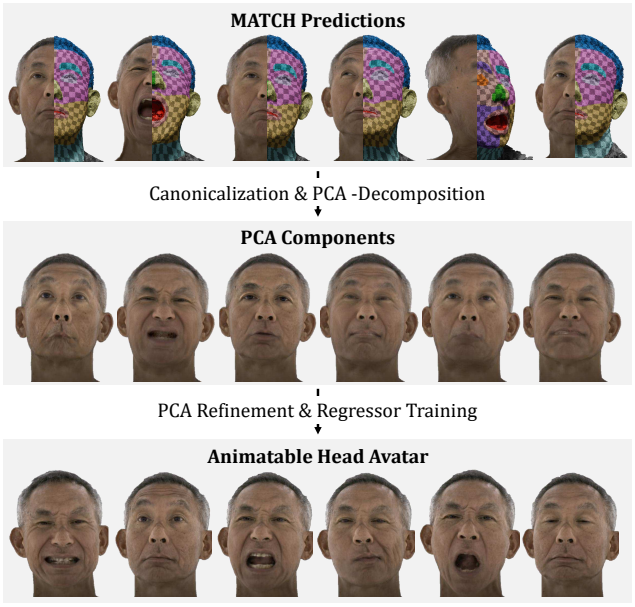


Figure 21. Procedure to create subject-specific head avatars from a sequence of Gaussian splat textures predicted by MATCH.

we fix the colors, scales, and opacities of the Gaussians in the oral cavity to their mean calculated across all training frames.

iv) Mean Refinement: GEM only refines the PCA basis vectors $\mathbf{B}_i$ against the target images using photometric losses. We found it beneficial to also refine the PCA means $\boldsymbol{\mu}_i$ during that stage.

H.2. Detailed Baseline Description

We compare our avatars with the optimization-based methods GaussianAvatars [53], RGBAvatar [32], and GEM [70]. GaussianAvatars optimizes Gaussian splats that are rigged to a parametric morphable face model against multi-view videos. RGBAvatar follows a similar approach, yet it also estimates Gaussian blendshapes from the face model parameters that can model dynamic appearance and geometry changes beyond the underlying face model. GEM first optimizes a CNN-based high-quality head avatar, which is then distilled into a lightweight, blendshape-based representation that can be directly animated from driving images.

GaussianAvatars and RGBAvatar can be directly driven with parameters of the FLAME 3DMM. For image-based animation, we estimate these parameters with EMOCA [8], a state-of-the-art 3DMM estimation method, which is also used as a pretrained feature extractor to drive GEM and our method. We found it beneficial for RGBAvatar to also use the EMOCA predictions during training. The performances for self- and cross-reenactment are evaluated on five subjects from the Ava-256 dataset. All methods are trained on a subset of the available sequences, avoiding extreme head and shoulder movements, protruded tongues, and isolated eye movements with a neutral face, while leaving out the `EXP_free_face` sequence for validation. Since we only aim to extract facial expressions from the driving image, not the global rigid transformation, we use the ground truth global pose from the VHAP tracking for the baselines and from MATCH's registrations for our method during evaluation.

H.3. Detailed Reconstruction Time Analysis

Table 10 presents a detailed breakdown of the time cost distribution across the individual stages of head avatar reconstruction for each method. The measurements were taken on a representative training sequence with 3212 frames using a compute node with a single NVIDIA A100 40GB GPU, 16 CPUs, and 500GB of RAM. To ensure full GPU usage during VHAP tracking, we ran two processes in parallel. File system operations, e.g., data loading and writing, were excluded from all timing computations since they are highly system-dependent. We find that the major bottleneck of the baseline's reconstruction time, especially for RGBAvatar, lies in the multi-view head tracking. While RGBAvatar reports an impressive reconstruction time of only $80s$ for the monocular setting, in the multi-view setting, they require optimization-based tracking with VHAP [52], which takes 10.65h on a representative 3212 frame training sequence, while the avatar optimization time increases to 0.75h[1]. GEM's multi-stage approach of first tracking a parametric head model, then optimizing a high-quality head avatar, followed by a distillation, even increases the total reconstruction time per avatar to 45.3h in our setup. Instead, MATCH allows for skipping the lengthy optimization-based mesh registration by directly predicting registered Gaussians from the multi-view images, which takes 0.53h for the entire training sequence compared to 10.65h of optimization-based tracking with VHAP. Unposing and PCA decomposition take 0.16 hours, such that we can start the refinement of the blendshapes and training of the expression regressor even before any of the baselines has completed registering just one 10th of the training frames.

H.4. Further qualitative comparisons

Figure 26 and Figure 27 present further results of the personalized head avatars for self- and cross-reenactment respectively.

H.5. Ablation Study

Figure 22 and Table 11 present qualitative and quantitative ablation studies of the changes applied to GEM [70]

[1]Experiments conducted with hyperparameters from the official code base.

Method	GA [53]	RGBAvatar [32]	GEM [70]	Ours
Stage-Wise Durations	VHAP Tracking: 10.65h Avatar Optimization: 4.83h	VHAP Tracking: 10.65h Avatar Optimization: 0.75h	VHAP Tracking: 10.65h CNN-Avatar Optimization: 27.70h Regressor Training: 6.94h	Coarse Mesh Registration: 0.09h MATCH Inference: 0.44h Canonicalization & PCA Decomp.: 0.16h Emoca & Deca Inference: 0.05h PCA Refinement: 2.75h Expression Regressor Training: 1.14h
Total Reconstruction Time	15.48h	11.40h	45.29h	4.63 h

Table 10. Head avatar reconstruction time breakdown. The measurements were conducted on a representative training sequence with 3,212 frames.

Modality-Specific PCA | Dynamic Mouth | 50 PCA Comp. | 100 PCA Comp. | Ours | Ground Truth

Figure 22. Qualitative ablation study of the changes applied to GEM [70] to create subject-specific head avatars from MATCH's predictions. Ours uses 150 PCA components.

	Self-Reenactment					Cross-Reenactment	
	LPIPS ↓	CSIM ↑	SSIM ↑	L1 ↓	PSNR ↑	CSIM ↑	EmoL1 ↓
Modality-Specific PCAs	0.180	0.879	0.816	0.027	24.376	0.815	9.849
Dynamic Mouth	0.174	0.878	0.809	0.027	24.112	0.811	9.792
# PCA Comp. = 50	0.177	0.876	0.808	0.027	24.126	0.814	9.891
# PCA Comp. = 100	0.175	0.880	0.809	0.027	24.112	0.814	9.907
Ours	0.174	0.880	0.809	0.027	24.122	0.813	9.837

Table 11. Quantitative ablation study of the changes applied to GEM [70] to create subject-specific head avatars from MATCH's predictions. By default, we use 150 PCA components.

to create subject-specific head avatars from MATCH's predictions. We find that employing separate PCAs for the individual Gaussian modalities ('Modality-Specific PCAs'), i.e., location, color, scale, rotation, and opacity, yields inferior results compared to jointly modeling all attributes in a single PCA. This aligns with the intuition that the different Gaussian attributes are highly correlated (e.g., raising the eyebrows results in darker colors for wrinkles on the forehead). Modeling the mouth interior with Gaussians with dynamically changing color, opacity, and scale ('Dynamic Mouth') does not change the quantitative scores significantly, yet slightly reduces the faithfulness of extreme expressions at test time (see Figure 22). We deduce that the lightweight image-based expression encoder fails to learn the intricate dynamics of the highly dynamic mouth interior Gaussians and benefits from additional consistency constraints enforced through static colors, opacities, and scales in this region. Increasing the number of PCA components improves the perceptual quality by adding high-frequency details. Note that even with the highest number of PCA components that we test, i.e., our default value of 150, we still use fewer components than GEM's modality-specific PCAs with a total of 180 components.

I. Ethical Considerations

Our method relies on multi-view studio captures with calibrated cameras, ensuring that all participants were aware of and consented to data collection. However, with the emergence of generative multi-view models such as CAP4D [58], similar data could be fabricated synthetically. This raises potential ethical concerns regarding consent and misuse, which we strongly discourage.

Figure 23. Additional novel view synthesis results on Ava-256 [45].

Figure 24. Additional novel view synthesis results on NeRSemble [28]. Ours (Ava) / Ours (NeRSemble) are trained on Ava-256 [45] and NeRSemble only, respectively.

Figure 25. Additional interpolation results. γ denotes the interpolation factor.

Figure 26. Additional self-reenactment results for the personalized head avatars.

Figure 27. Additional cross-reenactment results for the personalized head avatars.