Face Anything: 4D Face Reconstruction from Any Image Sequence

Umut Kocasarı[1], Simon Giebenhain[1], Richard Shaw[2], and Matthias Nießner[1]

[1] TU Munich
[2] Huawei Noah's Ark Lab

Fig. 1: Face Anything. Unified 4D facial reconstruction and dense tracking from image sequences via joint prediction of depth and canonical facial coordinates. Left to right: RGB input, 4D reconstruction with tracks, canonical maps, depth maps, and normal maps. Website: `https://kocasariumut.github.io/FaceAnything/`

Abstract. Accurate reconstruction and tracking of dynamic human faces from image sequences is challenging because non-rigid deformations, expression changes, and viewpoint variations occur simultaneously, creating significant ambiguity in geometry and correspondence estimation. We present a unified method for high-fidelity 4D facial reconstruction based on canonical facial point prediction, a representation that assigns each pixel a normalized facial coordinate in a shared canonical space. This formulation transforms dense tracking and dynamic reconstruction into a canonical reconstruction problem, enabling temporally consistent geometry and reliable correspondences within a single feed-forward model. By jointly predicting depth and canonical coordinates, our method enables accurate depth estimation, temporally stable reconstruction, dense 3D geometry, and robust facial point tracking within a single architecture. We implement this formulation using a transformer-based model that jointly predicts depth and canonical facial coordinates, trained using multi-view geometry data that non-rigidly warps into the canonical space. Extensive experiments on image and video benchmarks demonstrate state-of-the-art performance across reconstruction and tracking tasks, achieving approximately $3\times$ lower correspondence error and faster inference than prior dynamic reconstruction methods, while improving depth accuracy by 16%. These results highlight canonical facial point prediction as an effective foundation for unified feed-forward 4D facial reconstruction.

Keywords: 4D Face Reconstruction · Dynamic 3D Face Reconstruction · 3D Face Modeling · Dense / Sparse Point Tracking

1 Introduction

Reconstructing and tracking dynamic human faces from image sequences is a fundamental problem in computer vision with applications in virtual avatars, telepresence, animation, and human–computer interaction. Systems must recover detailed dynamic geometry and establish consistent correspondences across time from unconstrained image sequences. Despite recent progress in feed-forward 3D reconstruction [33, 62, 67], achieving temporally consistent 4D facial reconstruction with reliable tracking remains challenging. Faces exhibit complex nonrigid deformations caused by expressions and head motion while preserving fine geometric structures such as wrinkles, hair, and mouth interiors. Existing approaches often struggle to maintain both geometric fidelity and consistent correspondences under large expressions, extreme viewpoints, and long sequences.

Recovering dynamic facial geometry from image observations is fundamentally underconstrained and therefore requires strong learned priors. In addition to reconstructing geometry, 4D facial understanding requires correspondences across time, introducing an additional representational challenge. Most existing methods formulate tracking as predicting how points move across frames [15, 58, 59], requiring reasoning about mappings between frame pairs. Such motion-based formulations become increasingly difficult to learn as motion complexity and sequence length increase and require large amounts of supervision to obtain stable correspondences. In contrast, representing faces in a normalized canonical space is simpler, since canonical geometry is largely shared across poses and expressions and provides a stable reference for correspondence estimation.

Existing approaches address these challenges using different forms of geometric priors. Learning-based reconstruction models such as DA3 [33] learn strong geometric priors from large-scale data and achieve accurate monocular and multi-view depth prediction, but do not establish correspondences required for tracking. Face-specific predictors such as DAViD [53] and Sapiens [26] provide accurate single-image facial geometry estimation but operate independently per frame and therefore lack temporal consistency. Correspondence-based methods such as P3DMM [18] and V-DPM [58] estimate tracking, but parametric representations limit geometric detail while motion-based formulations require multiple forward passes and high computational cost. As a result, existing approaches treat reconstruction and correspondence estimation as separate problems or rely on representations that do not scale well to detailed dynamic faces.

In this work, we introduce a unified method for high-fidelity facial reconstruction and tracking based on canonical map prediction. Instead of predicting frame-to-frame motion, our method predicts a dense canonical map assigning each pixel a canonical facial coordinate in a normalized pose and expression space. Correspondences are obtained through nearest-neighbor search in canonical space, transforming tracking into a canonical reconstruction problem. This representation naturally enforces temporal consistency while remaining efficient to compute and provides normalized geometry suitable for downstream tasks such as animation and avatar generation. We implement the proposed formulation with a transformer-based architecture that jointly predicts depth and canonical facial

points. The network uses a DPT-style head [51] to process multiple input images simultaneously. Because learning canonical representations requires dense supervision that is largely absent from existing datasets, we construct a dataset based on NeRSemble [27] with high-quality multi-view reconstructions and canonical correspondences aligned using FLAME [30]. Extensive experiments demonstrate state-of-the-art performance in face depth estimation, temporally stable video depth prediction, dense 4D reconstruction, and facial point tracking. Compared with prior approaches such as V-DPM [58], our method achieves superior reconstruction and tracking accuracy while requiring less computation and memory.

In summary, this paper makes the following contributions:

- We propose a novel transformer-based method for unified 4D facial reconstruction and tracking. Unique properties of the face domain enable us to exploit canonical position map prediction, in addition to depth and ray maps, allowing for temporally stable correspondences.
- To supervise our novel formulation, we introduce a large-scale dynamic 3D face dataset, including canonicalized representations, obtained from high-quality MVS reconstructions and FLAME tracking.
- We achieve state-of-the-art performance on facial single-image and monocular video depth estimation, dense 4D reconstruction, and 3D point tracking.

2 Related Work

Static 3D and Feed-Forward Reconstruction. Classical 3D reconstruction relies on structure-from-motion and multi-view stereo pipelines that estimate cameras and dense geometry through global optimization [54, 55], with systems such as COLMAP [54] remaining standard for static scenes. Neural implicit representations such as NeRF and its extensions [1,6,39,42,45,46,61], together with explicit formulations including 3D Gaussian Splatting [7, 25, 38, 41], enable reconstruction via differentiable rendering but require scene-specific optimization.

Recent work focuses on feed-forward geometric inference from images, including correspondence-based reconstruction [28,66], unified multi-view transformer models [24,62,67], feed-forward splatting approaches [5,8,21,44,72,77], and large-scale geometry foundation models such as Depth Anything [33,75,76]. While enabling accurate feed-forward reconstruction, these approaches do not explicitly predict dense correspondences required for tracking deformable objects.

Dynamic 4D Reconstruction and Tracking. Dynamic extensions to neural implicit representations model non-rigid motion through time-conditioning and deformation fields [47,48,50], while dynamic Gaussian representations enable efficient time-dependent scene modeling and primitive tracking [20,31,37,57,68].

Recent work moves toward feed-forward dynamic reconstruction, including dynamic splatting and online reconstruction methods [32, 70, 74] and multi-view geometric frameworks predicting temporally aligned geometry or point maps across frames [22, 35, 65, 67, 78, 82]. Dynamic Point Map representations recover scene motion in a feed-forward manner [15, 58, 59]. Tracking-any-point methods estimate correspondences purely in image space [3, 10, 14, 23, 29]. While dynamic

reconstruction methods recover geometry and tracking approaches estimate correspondences, neither explicitly predicts dense canonical coordinates that provide identity-consistent alignment under non-rigid deformation.

Monocular Face Reconstruction and Parametric Models. Monocular face reconstruction typically relies on parametric 3D Morphable Models (3DMMs), such as the Basel Face Model and FLAME [17, 30, 49], representing identity and expression in low-dimensional subspaces. Optimization- and learning-based methods fit model parameters directly to images or video [4, 12, 52, 60]. Subsequent approaches further improve geometric detail and correspondence estimation [11, 16, 18, 43, 64] but remain constrained by fixed parametric topology or limited temporal modeling. Recent large-scale approaches such as DAViD [53] and Sapiens [26] improve single-image facial geometry prediction but do not enforce temporal consistency or dense correspondences.

Neural head avatar methods IMavatar [81] and Neural Head Avatars [19] reconstruct subject-specific dynamic heads via optimization. Splatting–based methods [9, 13, 56, 63, 69, 73, 79, 83] represent animatable heads using Gaussians for high-fidelity real-time rendering. However, these approaches rely on parametric representations or subject-specific optimization, leaving geometry and correspondence estimation largely decoupled. In contrast, our method predicts dense canonical coordinates supervised by parametric alignment but represented non-parametrically. This reduces correspondence estimation to reconstruction in canonical space, enabling unified feed-forward 4D reconstruction and tracking.

3 Method

3.1 Architecture

We reconstruct dynamic faces and establish dense correspondences using a single feed-forward network, as illustrated in Fig. 2. Instead of predicting frame-to-frame motion or deformation fields, we formulate correspondence estimation as **canonical map prediction**, where each pixel is mapped to a canonical facial coordinate. This formulation requires only one forward pass followed by efficient canonical-space matching, making it more efficient than motion-based approaches that require multiple evaluations. It also simplifies learning, since canonical geometry is normalized and structurally similar across poses and expressions, whereas deformation targets vary significantly with viewpoint and motion.

Our method employs a transformer-based architecture that predicts depth, ray maps, and canonical maps from one or more input images. The design is related to DA3 [33], which predicts depth and ray maps. To support canonical prediction, we incorporate a DPT head [51]. Given input images $\mathcal{I} = \{I_i\}_{i=1}^{N}$, the network predicts depth maps $D_i \in \mathbb{R}^{H\times W}$, ray maps $R_i \in \mathbb{R}^{H\times W\times 3}$, and canonical maps $C_i \in \mathbb{R}^{H\times W\times 3}$:

$$(D_i, R_i, C_i) = f_\theta(I_1, \ldots, I_N). \tag{1}$$

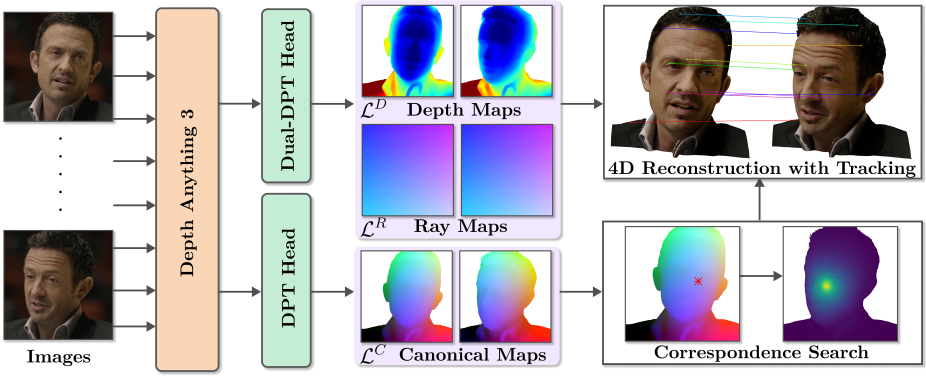

Fig. 2: Architecture overview. Given image sequences, our method jointly predicts depth and canonical facial maps to enable dense 4D reconstruction and tracking. Dense correspondences are established in canonical space, producing temporally consistent geometry and point trajectories.

Training. We train the model in two stages. First, the architecture is pretrained on DAViD [53] using monocular input to learn facial geometric priors, where depth is supervised using a masked L_1 regression loss. After pretraining, we add the canonical prediction head and finetune the full network on the dataset described in Sec. 3.3. The network predicts depth, ray, and canonical maps jointly.

To reduce the domain gap between multi-view capture and monocular video, training alternates between two sampling strategies: (1) multi-view images from a single timestamp and (2) single-camera images across multiple timestamps. This improves generalization to both reconstruction and tracking scenarios.

Correspondence Estimation. Dense correspondences are obtained by nearest-neighbor search in canonical space. Given images I_i and I_j, a pixel $\mathbf{p}$ in I_i corresponds to

$$\mathbf{q} = \arg\min_{\mathbf{q}' \in \Omega_j} \|C_i(\mathbf{p}) - C_j(\mathbf{q}')\|_2 , \tag{2}$$

where $\mathbf{q}$ denotes the corresponding pixel in image I_j and Ω_j is the set of pixels in I_j.

Nearest-neighbor search is implemented using KD-Tree [2] implementation and typically requires less than 0.2 seconds per image pair on a single CPU. The process is fully parallelizable across timestamps and can be accelerated via spatial downsampling with negligible accuracy loss.

Compared to deformation-based approaches such as V-DPM [58], canonical map prediction is both easier to learn and more efficient, producing correspondences in a single forward pass while maintaining high reconstruction fidelity.

3.2 Loss Formulation

For each prediction type $X \in \{D, R, C\}$ with ground truth X^*, we use regression, confidence-weighted regression, and gradient losses:

$$\mathcal{L}^X_{\text{reg}} = \frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} |X(\mathbf{p}) - X^*(\mathbf{p})| \tag{3}$$

$$\mathcal{L}^X_{\text{conf}} = \frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} (\gamma |X(\mathbf{p}) - X^*(\mathbf{p})| W_X(\mathbf{p}) - \alpha \log W_X(\mathbf{p})) \tag{4}$$

$$\mathcal{L}^X_{\text{grad}} = \frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} \left(|\nabla_x E_X(\mathbf{p})| + |\nabla_y E_X(\mathbf{p})| \right), \quad E_X = X - X^* \tag{5}$$

$$\mathcal{L} = \sum_{X \in \{D,R\}} \left(\mathcal{L}^X_{\text{reg}} + \mathcal{L}^X_{\text{conf}} + \mathcal{L}^X_{\text{grad}} \right) + \lambda_C \left(\mathcal{L}^C_{\text{reg}} + \mathcal{L}^C_{\text{conf}} + \mathcal{L}^C_{\text{grad}} \right) \tag{6}$$

where Ω denotes valid pixels, $W_X(\mathbf{p})$ is the predicted confidence, $\alpha = 0.2$ and $\gamma = 1$ are scalar weighting parameters, and $\lambda_C = 5$ weights the canonical map losses.

3.3 Dataset Creation

Learning canonical facial correspondences requires supervision that is largely absent from existing datasets. To address this limitation, we construct a new dataset based on the NeRSemble dataset [27], which provides synchronized multi-view videos with calibrated cameras. We first pretrain our model on the DAViD dataset [53] to learn strong human-specific priors, and then finetune on the NeRSemble-based dataset to learn detailed facial geometry and canonical correspondences. NeRSemble contains multi-view recordings captured with 16 cameras across a wide range of subjects, poses, and expressions. From this dataset, we use 414 subjects and approximately 20k timestamps across multiple sequences, corresponding to roughly 320k images.

Rather than sampling timestamps uniformly, we select frames that maximize diversity in facial expressions and head poses. To achieve this, we run MediaPipe [36] on the frontal camera view to estimate facial blendshape parameters and head poses for all sequences. For each subject, we select 50 timestamps using farthest point sampling to ensure diverse coverage. We separately sample 40 timestamps based on blendshape parameters and 10 timestamps based on pose parameters in order to balance expression and pose variation.

For each selected timestamp, we reconstruct geometry using COLMAP [54] from all 16 camera views. Reconstruction is performed on images downsampled by a factor of four for computational efficiency. We retain only reconstructions where all views are successfully registered, as incomplete registrations often indicate unreliable geometry. The resulting reconstructions provide multi-view consistent depth maps and dense point clouds with detailed facial geometry. To improve reconstruction quality in challenging regions such as hair, we adjust COLMAP hyperparameters to obtain more complete point coverage.

To establish canonical correspondences, we perform FLAME-based tracking [30] for each selected timestamp. For each subject, we enforce consistent

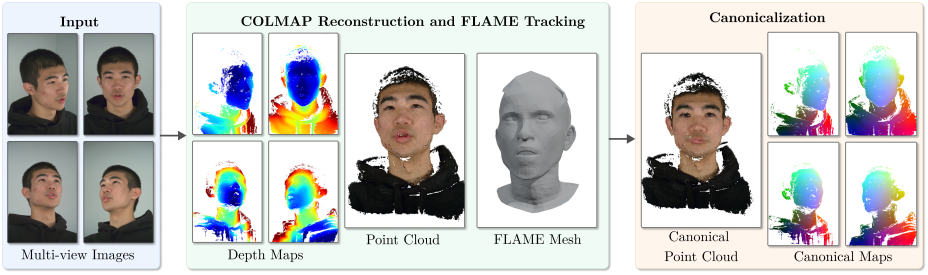

Fig. 3: Dataset creation. We generate training supervision by combining multi-view reconstruction with parametric face tracking to produce depth maps and canonical facial maps. Although the parametric face model may not capture fine-scale geometric details, high-frequency information from COLMAP reconstruction is preserved in the canonical maps. This process provides geometrically consistent supervision across viewpoints, expressions, and identities for training our model.

shape and static offset parameters across all timestamps and remove outliers based on tracking error statistics. This produces stable alignments between reconstructed geometry and parametric face models.

Our goal is to preserve the geometric detail of COLMAP reconstructions while aligning them into a shared canonical space. We construct canonical point clouds by transferring deformations estimated from FLAME tracking to reconstructed points. Given tracked and canonical FLAME meshes, we compute per-vertex deformation vectors between corresponding surface points. Each COLMAP point is assigned the deformation of its nearest FLAME surface point, allowing reconstructed points to be mapped into canonical FLAME coordinate system.

Using these canonicalized point clouds, we generate canonical maps that describe the canonical location of each pixel in the input images. These maps provide dense supervision for learning canonical correspondences, while the COLMAP depth maps provide supervision for detailed geometry reconstruction. Although depth supervision can be sparse in some regions, the strong geometric prior learned from DA3 enables reliable generalization to these areas.

The resulting dataset provides multi-view RGB images, calibrated camera parameters, depth maps, and dense canonical maps across a wide range of identities, poses, and expressions. Because canonical maps are defined in a normalized FLAME coordinate system, correspondences across subjects and frames exhibit strong structural consistency, making them well-suited for supervised learning. Despite minor inaccuracies from parametric tracking, we observe that the network learns to compensate for small alignment errors during training. An overview of the dataset creation process is shown in Fig. 3.

4 Experiments

4.1 Implementation Details

We train the model in two stages. First, our 1.2B-parameter model is pretrained on approximately 100k facial images from the DAViD dataset [53] to learn ge-

Fig. 4: 4D reconstruction comparison on VFHQ, NeRSemble, and Ava-256. COLMAP reconstructions are shown as pseudo ground truth for NeRSemble and Ava-256, while VFHQ does not provide ground-truth geometry. Our method produces more accurate and detailed reconstructions than recent approaches.

ometric facial priors. Then, the main training stage is performed for 90 epochs using the AdamW optimizer [34] with a learning rate initialized at 2×10^{-5} and decayed to 1×10^{-8} using a cosine scheduler, together with gradient clipping of 1. Each epoch consists of 800 sampled batches containing up to 48 images, while each sequence contains between 2 and 16 images with gradient accumulation over three steps. Training is performed using bfloat16 precision with gradient checkpointing for efficiency. Following DA3 [33], we train using multiple input resolutions with a base resolution of 504×504.

4.2 Evaluation Protocol

Evaluation Datasets. We evaluate our method on various datasets to measure both reconstruction fidelity and generalization to in-the-wild data, including NeRSemble [27], Ava-256 [40], VFHQ [71], and CelebV-HQ [84]. From NeRSemble and Ava-256 we select five subjects each; NeRSemble subjects are not used during training and include diverse sequences, while Ava-256 samples are obtained by randomly selecting 40 images across all sequences. For VFHQ, we evaluate on all test videos, and for CelebV-HQ, we randomly select a subset of videos. For consistency across datasets, we use the frontal camera views.

Evaluation Metrics. Depth accuracy is evaluated using the Root Mean Squared Error (RMSE) and Absolute Relative Error (AbsRel). Correspondence

Table 1: Monocular depth estimation on NeRSemble and Ava-256. Results are reported for both images and videos in metric scale as Image/Video. Best results are shown in bold. Values are $\times 10$.

	NeRSemble		Ava-256	
Setting	RMSE↓	AbsRel↓	RMSE↓	AbsRel↓
Pi3 [67]	0.193/0.160	0.134/0.108	0.083/0.091	0.066/0.071
DA3 [33]	0.162/0.127	0.100/0.085	0.148/0.116	0.129/0.100
VGGT [62]	0.115/0.119	0.076/0.080	0.084/0.103	0.061/0.084
Sapiens-1B [26]	0.112/0.112	0.065/0.065	0.079/0.079	0.059/0.059
DAVID [53]	0.110/0.110	0.061/0.061	0.182/0.182	0.160/0.160
V-DPM [58]	0.102/0.104	0.061/0.062	0.090/0.097	0.070/0.075
Sapiens-2B [26]	0.085/0.085	0.048/0.048	0.081/0.081	0.056/0.056
Ours	**0.077/0.075**	**0.040/0.038**	**0.067/0.065**	**0.048/0.048**

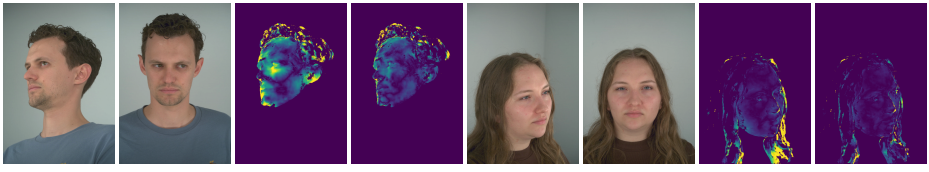

Fig. 5: Single-view vs multi-view depth prediction. Darker colors indicate lower error. Multi-view input improves depth accuracy over monocular prediction.

accuracy is measured using end-point error (EPE) in both 2D and 3D, where $P_i(t_j)$ and various temporal margins follow the definition in V-DPM [58]. We additionally report the percentage of correspondences within different pixel thresholds ($< n$px). Geometric consistency is evaluated using Forward-Backward Cycle Consistency Error (CCE) of correspondence predictions, while photometric alignment is assessed with Warping Photometric Error (WPE) after warping correspondences between frames. FLAME tracking accuracy is measured using the unidirectional L1 Chamfer distance (CD-L1) from ground-truth COLMAP points to the predicted mesh.

4.3 4D Reconstruction

We evaluate our method on 4D facial reconstruction and compare against state-of-the-art approaches in qualitative and quantitative settings. For multi-image methods, we provide all input views and evaluate on the first 40 frames of each sequence, while monocular methods are applied independently to each frame.

Qualitative Reconstruction. Fig. 4 presents qualitative comparisons of reconstructed 3D facial geometry across methods. Our approach produces high-fidelity reconstructions with detailed facial geometry, whereas baseline methods often exhibit geometric artifacts or reduced detail.

Fig. 6: 2D tracking comparison on VFHQ. Track points are defined in the base image and each method predicts trajectories to the target image that should end at the same facial locations. Our method produces more accurate and consistent correspondences than recent approaches.

Depth Accuracy. Depth evaluation is reported in Tab. 1 for both image-based and video-based reconstruction settings. In the video-based setting, all frames of the input sequence are provided jointly as input. Our method achieves state-of-the-art performance across datasets despite using a smaller model than the largest Sapiens [26] variant and requiring significantly less training time.

Effect of Multi-Image Prediction. To analyze the impact of multi-image inference on the NeRSemble dataset, Fig. 5 compares depth error maps obtained using a single input image and using an additional supporting view. Incorporating supporting images clearly improves depth accuracy, particularly in side-view regions where monocular prediction is more ambiguous.

4.4 2D Tracking

We evaluate dense 2D tracking accuracy on both controlled and in-the-wild datasets and compare against state-of-the-art facial correspondence methods.

Qualitative Tracking. Fig. 6 shows qualitative comparisons of 2D correspondences on the VFHQ [71] dataset. We visualize tracks between a reference image and a target frame for different methods. Our approach produces more consistent and accurate tracks, particularly in regions with large motion and fine structures. P3DMM [18] fails to track points reliably in hair regions, while

Table 2: 2D correspondence evaluation on NeRSemble. Dense correspondence accuracy is measured using EPE and percentage of predictions within pixel thresholds across temporal margins.

	Margin = 2				Margin = 8			
Method	EPE(2D)↓	<3px↑	<5px↑	<10px↑	EPE(2D)↓	<3px↑	<5px↑	<10px↑
P3DMM [18] (w/o Hair)	3.089	0.629	0.873	0.986	3.971	0.508	0.768	0.955
Ours (w/o Hair)	**1.719**	**0.876**	**0.968**	**0.995**	**2.314**	**0.769**	**0.926**	**0.990**
P3DMM [18] (w/ Hair)	5.550	0.565	0.797	0.927	6.597	0.446	0.686	0.883
Ours (w/ Hair)	**1.838**	**0.853**	**0.960**	**0.995**	**2.470**	**0.741**	**0.913**	**0.987**

Table 3: Dense correspondence evaluation on VFHQ. Correspondence accuracy is evaluated using CCE and WPE at different temporal margins.

	Margin = 5					
Method	$CCE_{mean}\downarrow$	$CCE_{median}\downarrow$	$CCE_{<2px}\uparrow$	$WPE_{L1}\downarrow$	$WPE_{Grad}\downarrow$	$WPE_{SSIM}\uparrow$
P3DMM [18]	2.472	1.146	0.702	0.032	0.016	0.798
V-DPM [58]	0.797	0.610	0.885	0.026	0.015	0.855
Ours	**0.398**	**0.007**	**0.961**	**0.023**	**0.014**	**0.881**
	Margin = 20					
Method	$CCE_{mean}\downarrow$	$CCE_{median}\downarrow$	$CCE_{<2px}\uparrow$	$WPE_{L1}\downarrow$	$WPE_{Grad}\downarrow$	$WPE_{SSIM}\uparrow$
P3DMM [18]	2.797	1.146	0.673	0.043	0.017	0.740
V-DPM [58]	1.348	1.054	0.756	0.040	0.017	0.769
Ours	**0.774**	**0.069**	**0.909**	**0.034**	**0.016**	**0.810**

V-DPM [58] frequently predicts inaccurate locations under large motions. In contrast, our method maintains stable correspondences across challenging frames.

Quantitative Correspondence Accuracy. We report quantitative tracking accuracy on the NeRSemble [27] dataset using ground-truth correspondences in Tab. 2. Our method consistently outperforms P3DMM across different temporal intervals and for both face-only and full head regions including hair.

Cycle and Photometric Consistency. Additional quantitative results on the VFHQ dataset are reported in Tab. 3, where we compare against P3DMM and V-DPM using CCE and WPE. Our method achieves the best performance.

4.5 Point Tracking

We evaluate dense 3D point tracking accuracy and efficiency of our method.

Qualitative Tracking. Fig. 7 presents qualitative comparisons with V-DPM on the CelebV-HQ [84] dataset. Our method produces more accurate 4D reconstructions and more stable point trajectories over time, while V-DPM exhibits larger geometric inconsistencies and tracking errors. The results demonstrate that our canonical-space formulation leads to more reliable correspondences.

Quantitative Tracking Accuracy. Tab. 4 reports tracking accuracy across temporal margins. Our method achieves lower errors than V-DPM for both short- and long-range correspondences.

Efficiency Analysis. We compare inference efficiency in Tab. 5, reporting runtime and GPU memory usage measured on 40 images and the maximum

Table 4: 3D correspondence evaluation on NeRSemble. 3D correspondence accuracy is measured using EPE across temporal margins for reconstruction and tracking.

	Margin = 2				Margin = 8				Margin = 1–10
Method	$P_0(t_0)\downarrow$	$P_0(t_1)\downarrow$	$P_1(t_0)\downarrow$	$P_1(t_1)\downarrow$	$P_0(t_0)\downarrow$	$P_0(t_1)\downarrow$	$P_1(t_0)\downarrow$	$P_1(t_1)\downarrow$	EPE$\downarrow$
V-DPM [58]	0.014	0.015	0.014	0.014	0.014	0.016	0.014	0.015	0.015
Ours	**0.004**	**0.004**	**0.004**	**0.004**	**0.004**	**0.005**	**0.005**	**0.004**	**0.005**

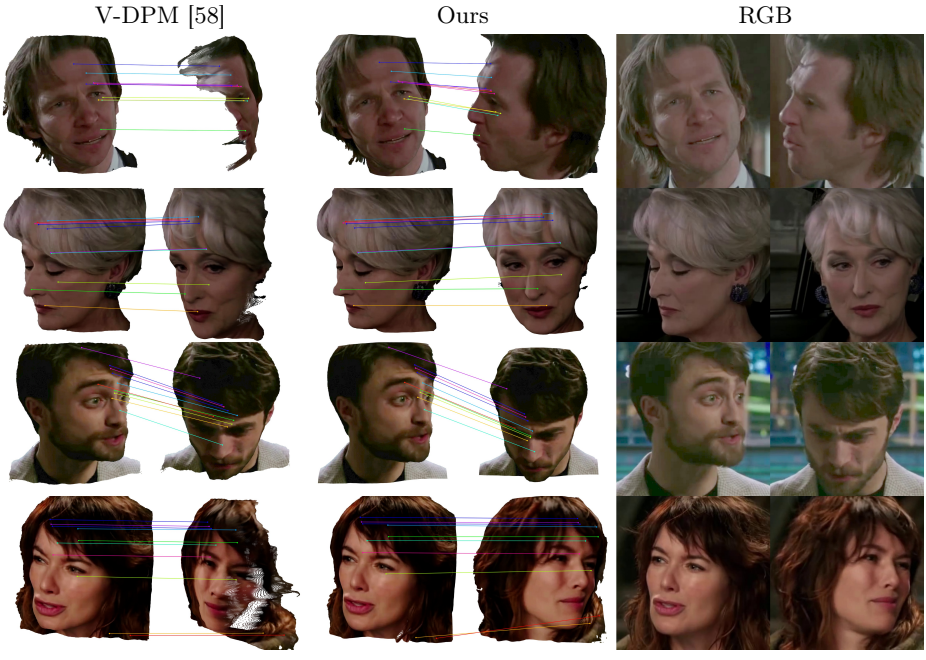

Fig. 7: 4D reconstruction and tracking comparison on CelebV-HQ. Both input images are reconstructed and correspondences are visualized with track lines between frames. Camera locations are fixed in the first image across methods for fair comparison.

number of images that fit on a single GPU (all experiments at 518×518). Our method achieves favorable efficiency while supporting larger batch sizes.

Temporal Correspondence Accuracy. Fig. 8 shows correspondence error maps across temporal intervals. Our method produces consistently lower errors and cleaner correspondence structures compared to competing approaches, further demonstrating the accuracy of our tracking formulation.

4.6 FLAME Tracking

We evaluate monocular FLAME tracking accuracy when constrained by predictions from P3DMM and our method. As shown in Tab. 6, our predictions lead to more accurate face tracking than constraints derived from P3DMM.

Table 5: Runtime and memory efficiency comparison. Inference runtime and GPU memory usage measured on a single GPU (80GB).

Method	Runtime (s)↓	Peak Memory (GB)↓	Max Images/GPU↑
V-DPM [58]	160	40	74
Ours	**5**	**19**	**470**

Table 6: FLAME tracking accuracy on NeRSemble. Values are $\times 100$.

Method	CD-L1↓
P3DMM [18]	0.238
Ours	**0.195**

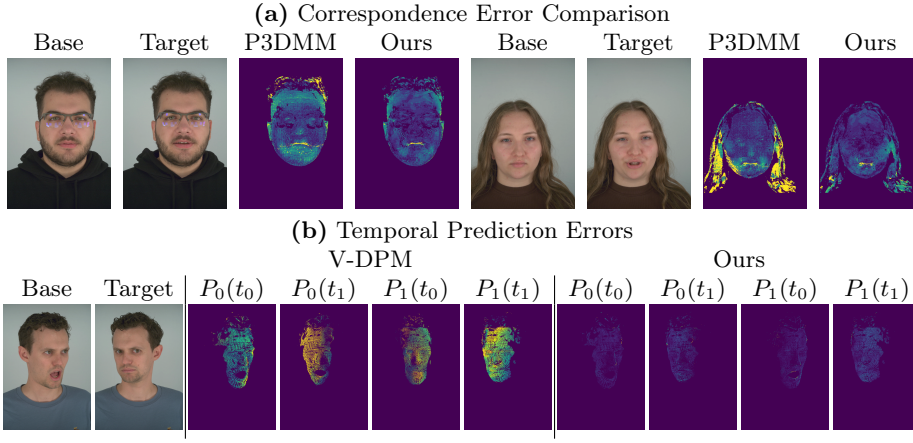

Fig. 8: Correspondence and temporal prediction errors on NeRSemble. (a) Pixel-wise correspondence errors for P3DMM [18] and our method. (b) Temporal prediction errors compared to V-DPM [58].

4.7 Ablation Studies

We conduct ablation studies on the NeRSemble dataset to analyze the effects of our training strategy and canonical correspondence formulation, as shown in Tab. 7. Starting from DA3 as the baseline, we observe limited performance, particularly for multi-view depth estimation. Pretraining on DAViD improves monocular depth accuracy thanks to stronger facial priors, but does not generalize well to multi-view depth prediction.

We further evaluate two simplified training variants. Monocular Training, where the camera viewpoint is fixed within a forward pass while the timestamp varies, yields strong correspondence and camera estimation accuracy but leads to weaker multi-view depth predictions. In contrast, Static Training, where the timestamp is fixed while the camera viewpoint varies, improves multi-view depth estimation but degrades correspondence and camera estimation due to the lack of temporal variation.

We also evaluate an alternative correspondence formulation (Ours (Motion Pred)) that predicts motion from the current timestamp to the canonical space. This formulation performs significantly worse in correspondence prediction, suggesting that directly learning motion mappings is less stable.

Our final formulation samples both cameras and timestamps during training and predicts canonical maps instead of motion fields. As shown in Tab. 7, this design achieves consistently strong performance across all metrics, producing

Table 7: Ablation study. Component analysis using depth (AbsRel), correspondence (EPE(2D)), and camera (Camera Rot (°)) metrics. Best results in **bold**, second-best underlined. AbsRel values are ×10. The full design achieves consistently strong performance across all metrics, producing accurate monocular and multi-view depth predictions while maintaining reliable correspondence estimation.

Method	AbsRel (Monocular)↓	EPE↓	AbsRel (16 views)↓	Camera Rot ↓
DA3 (Baseline)	0.085	-	0.076	-
DAViD Pretrained	0.061	-	0.033	-
Monocular Training	0.054	**3.031**	0.064	**0.038**
Static Training	**0.050**	3.864	**0.015**	2.102
Ours (Motion Pred)	-	6.210	-	-
Ours	0.053	3.271	**0.015**	0.054

accurate monocular and multi-view depth predictions while maintaining reliable correspondence estimation.

5 Limitations and Future Work

Despite achieving strong performance on 4D facial reconstruction and tracking, our method has several limitations. The model is specialized for faces and relies on learned facial priors, which limits generalization to non-face scenes, and canonicalization of nearby objects such as microphones, hands, or accessories is often unreliable. Reconstruction quality can also degrade under strong occlusions, extreme viewpoints, or limited facial visibility where geometric cues are insufficient. Extending canonical map prediction beyond faces and improving robustness in challenging real-world scenarios are promising directions for future work. Another promising direction is integrating canonical prediction with generative or neural rendering models to enable controllable facial animation and avatar creation from monocular video.

6 Conclusion

We present Face Anything, a unified method for high-fidelity 4D facial reconstruction and dense tracking from image sequences. Our method jointly estimates depth and canonical facial coordinates, enabling temporally consistent reconstruction and reliable correspondences within a single feed-forward model. The proposed canonical representation simplifies correspondence learning and allows efficient multi-frame reconstruction and tracking without explicit motion modeling. To support supervised learning of canonical correspondences, we construct a dataset based on NeRSemble with multi-view geometry and canonical alignment. Extensive experiments demonstrate state-of-the-art performance across face depth estimation, 4D reconstruction, and dense tracking benchmarks while improving efficiency over prior methods. These results highlight canonical map prediction as an effective representation for dynamic facial understanding and a promising direction for future spatiotemporal reconstruction methods.

Acknowledgments This work was supported by the ERC Consolidator Grant Gen3D (101171131) of Matthias Nießner. We thank Angela Dai for the video voice-over.

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., et al.: Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In: ICCV (2021)
2. Bentley, J.L.: Multidimensional Binary Search Trees used for Associative Searching. Communications of the ACM **18**(9), 509–517 (1975)
3. Bian, W., Huang, Z., Shi, X., Dong, Y., Li, Y., et al.: Context-PIPs: Persistent Independent Particles Demands Context Features. In: NeurIPS (2023)
4. Blanz, V., Vetter, T.: A Morphable Model for the Synthesis of 3D Faces. In: SIGGRAPH (1999)
5. Charatan, D., Li, S., Tagliasacchi, A., Sitzmann, V.: pixelSplat: 3D Gaussian Splats from Image Pairs for Scalable Generalizable 3D Reconstruction. In: CVPR (2024)
6. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: TensoRF: Tensorial Radiance Fields. In: ECCV (2022)
7. Chen, Y., Jiang, J., Jiang, K., Tang, X., Li, Z., et al.: DashGaussian: Optimizing 3D Gaussian Splatting in 200 Seconds. In: CVPR (2025)
8. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., et al.: MVSplat: Efficient 3D Gaussian Splatting from Sparse Multi-View Images. In: ECCV (2024)
9. Chen, Y., Wang, L., Li, Q., Xiao, H., Zhang, S., et al.: MonoGaussianAvatar: Monocular Gaussian Point-based Head Avatar. ACM Trans. Graph., Proc. SIGGRAPH (2024)
10. Cho, S., Huang, J., Nam, J., An, H., Kim, S., et al.: Local All-Pair Correspondence for Point Tracking. In: ECCV (2024)
11. Danecek, R., Black, M.J., Bolkart, T.: EMOCA: Emotion Driven Monocular Face Capture and Animation. In: CVPR (2022)
12. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., et al.: Accurate 3D Face Reconstruction With Weakly-Supervised Learning: From Single Image to Image Set. In: CVPRW (2019)
13. Dhamo, H., Nie, Y., Moreau, A., Song, J., Shaw, R., et al.: HeadGaS: Real-Time Animatable Head Avatars via 3D Gaussian Splatting. In: ECCV (2024)
14. Doersch, C., Yang, Y., Vecerik, M., Gokay, D., Gupta, A., et al.: TAPIR: Tracking Any Point with per-frame Initialization and Temporal Refinement. In: ICCV (2023)
15. Feng, H., Zhang, J., Wang, Q., Ye, Y., Yu, P., et al.: St4RTrack: Simultaneous 4D Reconstruction and Tracking in the World. In: ICCV (2025)
16. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network. In: ECCV (2018)
17. Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Lüthi, M., et al.: Morphable Face Models - An Open Framework. In: IEEE Conference on Automatic Face and Gesture Recognition. pp. 75–82 (2018)
18. Giebenhain, S., Kirschstein, T., Rünz, T., Agapito, L., Nießner, M.: Pixel3DMM: Versatile Screen-Space Priors for Single-Image 3D Face Reconstruction. arXiv preprint arXiv:2505.00615 (2025)
19. Grassal, P.W., Prinzler, M., Leistner, T., Rother, C., Nießner, M., et al.: Neural Head Avatars from Monocular RGB Videos. In: CVPR (2022)

20. Huang, Y.H., Sun, Y.T., Yang, Z., Lyu, X., Cao, Y.P., et al.: SC-GS: Sparse-Controlled Gaussian Splatting for Editable Dynamic Scenes. In: CVPR (2024)
21. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., et al.: AnySplat: Feed-forward 3d Gaussian Splatting from Unconstrained Views. ACM Trans. Graph. **44**(6), 1–16 (2025)
22. Jiang, Z., Zheng, C., Laina, I., Larlus, D., Vedaldi, A.: Geo4D: Leveraging Video Generators for Geometric 4D Scene Reconstruction. In: ICCV (2025)
23. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., et al.: CoTracker: It is Better to Track Together. In: ECCV (2024)
24. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., et al.: MapAnything: Universal Feed-Forward Metric 3D Reconstruction. In: 3DV (2026)
25. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graph. **42**(4) (July 2023)
26. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., et al.: Sapiens: Foundation for Human Vision Models. In: ECCV (2024)
27. Kirschstein, T., Qian, S., Giebenhain, S., Walter, T., Nießner, M.: NeRSemble: Multi-view Radiance Field Reconstruction of Human Heads. ACM Trans. Graph. **42**(4), 1–14 (2023)
28. Leroy, V., Cabon, Y., Revaud, J.: Grounding Image Matching in 3D with MASt3R. In: ECCV (2024)
29. Li, H., Zhang, H., Liu, S., Zeng, Z., Ren, T., et al.: TAPTR: Tracking Any Point with Transformers as Detection. In: ECCV (2024)
30. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a Model of Facial Shape and Expression from 4D Scans. ACM Trans. Graph., (Proc. SIGGRAPH Asia) **36**(6), 194:1–194:17 (2017)
31. Li, Z., Chen, Z., Li, Z., Xu, Y.: Spacetime Gaussian Feature Splatting for Real-Time Dynamic View Synthesis. In: CVPR (2024)
32. Lin, C., Lin, Y., Pan, P., Yu, Y., Hu, T., et al.: MoVieS: Motion-Aware 4D Dynamic View Synthesis in One Second. In: CVPR (2026)
33. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth Anything 3: Recovering the Visual Space from Any Views. arXiv preprint arXiv:2511.10647 (2025)
34. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: ICLR (2019)
35. Lu, J., Huang, T., Li, P., Dou, Z., Lin, C., et al.: Align3R: Aligned Monocular Depth Estimation for Dynamic Videos. In: CVPR (2025)
36. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., et al.: MediaPipe: A Framework for Building Perception Pipelines. arXiv preprint arXiv:1906.08172 (2019)
37. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. In: 3DV (2024)
38. Mallick, S.S., Goel, R., Kerbl, B., Steinberger, M., Carrasco, F.V., et al.: Taming 3DGS: High-Quality Radiance Fields with Limited Resources. In: SIGGRAPH Asia. Association for Computing Machinery, New York, NY, USA (2024)
39. Martin-Brualla, R., Radwan, N., Sajjadi, M.S.M., Barron, J.T., Dosovitskiy, A., et al.: NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In: CVPR (2021)
40. Martinez, J., Kim, E., Romero, J., Bagautdinov, T., Saito, S., et al.: Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. NeurIPS Track on Datasets and Benchmarks (2024)

41. Meuleman, A., Shah, I., Lanvin, A., Kerbl, B., Drettakis, G.: On-the-fly Reconstruction for Large-Scale Novel View Synthesis from Unposed Images. ACM Trans. Graph. **44**(4) (2025)
42. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., et al.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In: ECCV (2020)
43. Ming, X., Han, Y., Huang, T., Xu, F.: VGGTFace: Topologically Consistent Facial Geometry Reconstruction in the Wild. In: AAAI (2026)
44. Moreau, A., Shaw, R., Nazarczuk, M., Shin, J., Tanay, T., et al.: Off The Grid: Detection of Primitives for Feed-Forward 3D Gaussian Splatting. In: CVPR (2026)
45. Müller, T., Evans, A., Schied, C., Keller, A.: Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. ACM Trans. Graph. **41**(4), 102:1–102:15 (Jul 2022)
46. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S.M., Geiger, A., et al.: RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs. In: CVPR (2022)
47. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., et al.: Nerfies: Deformable Neural Radiance Fields. In: ICCV (2021)
48. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., et al.: HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields. ACM Trans. Graph. **40**(6) (dec 2021)
49. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3D Face Model for Pose and Illumination Invariant Face Recognition. In: IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments (2009)
50. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: Neural Radiance Fields for Dynamic Scenes. In: CVPR (2020)
51. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision Transformers for Dense Prediction. In: ICCV (2021)
52. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning Detailed Face Reconstruction from a Single Image. In: CVPR (2017)
53. Saleh, F., Aliakbarian, S., Hewitt, C., Petikam, L., Xiao, X., et al.: David: Data-efficient and Accurate Vision Models from Synthetic Data. In: ICCV (2025)
54. Schönberger, J.L., Frahm, J.M.: Structure-from-Motion Revisited. In: CVPR (2016)
55. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise View Selection for Unstructured Multi-View Stereo. In: ECCV (2016)
56. Shaw, R., Jang, Y., Papaioannou, A., Moreau, A., Dhamo, H., et al.: ICo3D: An Interactive Conversational 3D Virtual Human. IJCV (2025)
57. Shaw, R., Song, J., Moreau, A., Nazarczuk, M., Catley-Chandar, S., et al.: SWinGS: Sliding Windows for Dynamic 3D Gaussian Splatting. In: ECCV (2024)
58. Sucar, E., Insafutdinov, E., Lai, Z., Vedaldi, A.: V-DPM: 4D Video Reconstruction with Dynamic Point Maps. arXiv preprint arXiv:2601.09499 (2025)
59. Sucar, E., Lai, Z., Insafutdinov, E., Vedaldi, A.: Dynamic Point Maps: A Versatile Representation for Dynamic 3D Reconstruction. In: ICCV (2025)
60. Tewari, A.K., Zollhöfer, M., Kim, H., Garrido, P., Bernard, F., et al.: MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In: ICCV (2017)
61. Wang, G., Chen, Z., Loy, C.C., Liu, Z.: SparseNeRF: Distilling Depth Ranking for Few-shot Novel View Synthesis. In: ICCV (2023)

62. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., et al.: VGGT: Visual Geometry Grounded Transformer. In: CVPR (2025)
63. Wang, J., Xie, J.C., Li, X., Xu, F., Pun, C.M., et al.: GaussianHead: High-fidelity Head Avatars with Learnable Gaussian Derivation. IEEE Trans. on Visualization and Computer Graphics (2025)
64. Wang, L., Chen, Z., Yu, T., Ma, C., Li, L., Liu, Y.: FaceVerse: a Fine-grained and Detail-controllable 3D Face Morphable Model from a Hybrid Dataset. In: CVPR (June 2022)
65. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3D Perception Model with Persistent State. In: CVPR (2025)
66. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: Geometric 3D Vision Made Easy. In: CVPR (2024)
67. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., et al.: Pi3: Permutation-Equivariant Visual Geometry Learning. In: ICLR (2026)
68. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., et al.: 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In: CVPR (2024)
69. Wu, Z., Zhou, B., Hu, L., Liu, H., Sun, Y., et al.: UIKA: Fast Universal Head Avatar from Pose-Free Images. In: CVPR (2026)
70. Wu, Z., Yan, Q., Yi, X., Wang, L., Liao, R.: StreamSplat: Towards Online Dynamic 3D Reconstruction from Uncalibrated Video Streams. In: ICLR (2026)
71. Xie, L., Wang, X., Zhang, H., Dong, C., Shan, Y.: VFHQ: A High-Quality Dataset and Benchmark for Video Face Super-Resolution. In: CVPRW (2022)
72. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., et al.: DepthSplat: Connecting Gaussian Splatting and Depth. In: CVPR (2025)
73. Xu, Y., Chen, B., Li, Z., Zhang, H., Wang, L., et al.: Gaussian Head Avatar: Ultra High-fidelity Head Avatar via Dynamic Gaussians. In: CVPR (2024)
74. Xu, Z., Li, Z., Dong, Z., Zhou, X., Newcombe, R., et al.: 4DGT: Learning a 4D Gaussian Transformer Using Real-World Monocular Videos. In: NeurIPS (2025)
75. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In: CVPR (2024)
76. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything V2. arXiv preprint arXiv:2406.09414 (2024)
77. Ye, B., Liu, S., Xu, H., Xueting, L., Pollefeys, M., et al.: No Pose, No Problem: Surprisingly Simple 3D Gaussian Splats from Sparse Unposed Images. In: ICLR (2025)
78. Zhang, J., Herrmann, C., Hur, J., Jampani, V., Darrell, T., et al.: MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion. In: ICLR (2025)
79. Zhao, Z., Bao, Z., Li, Q., Qiu, G., Liu, K.: PSAvatar: A Point-based Shape Model for Real-Time Head Avatar Animation with 3D Gaussian Splatting. arXiv preprint arXiv:2401.12900 (2024)
80. Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., Wen, F.: General facial representation learning in a visual-linguistic manner. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18697–18709 (2022)
81. Zheng, Y., Abrevaya, V.F., Bühler, M.C., Chen, X., Black, M.J., Hilliges, O.: IM Avatar: Implicit Morphable Head Avatars from Videos. In: CVPR (2022)
82. Zhou, K., Zhou, K., Wang, Y., Chen, G., Beaudouin, G., et al.: PAGE-4D: Disentangled Pose and Geometry Estimation for 4D Perception. In: ICLR (2026)
83. Zhou, Z., Ma, F., Fan, H., Yang, Z., Yang, Y.: HeadStudio: Text to Animatable Head Avatars with 3D Gaussian Splatting. In: ECCV (2024)

84. Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., et al.: CelebV-HQ: A Large-Scale Video Facial Attributes Dataset. In: ECCV (2022)

Appendix

A Supplementary Video

We highly recommend watching our supplementary video, which presents additional qualitative results of our method. The video demonstrates **4D facial reconstruction with point tracking**, illustrating the temporal consistency of the reconstructed geometry across frames.

We further provide orbiting camera visualizations of the reconstructed sequences to better showcase the recovered dynamic facial geometry from novel viewpoints.

Additionally, the video includes comparisons of 2D correspondence tracking between our method and baseline approaches. These visualizations highlight the accuracy and temporal stability of the correspondences estimated by our approach.

B Additional Implementation Details

B.1 Training Details

During training, we apply a set of photometric augmentations to improve robustness to appearance variations. Specifically, we employ color jittering with brightness, contrast, and saturation factors of 0.5, and hue variation of 0.1. The color jitter augmentation is applied with probability 0.9. In addition, grayscale augmentation is enabled during training.

To maintain photometric consistency across correlated inputs, we apply *co-jittering*, where identical color perturbations are applied jointly across frames. This helps preserve relative color relationships between views while still providing appearance variation.

Training images are sampled from a predefined set of resolutions to improve robustness to scale changes. The set of resolutions used during training is:

- $(504, 504)$
- $(378, 504)$
- $(336, 504)$
- $(280, 504)$
- $(504, 336)$
- $(504, 756)$
- $(504, 672)$

For numerical stability, we normalize the ground-truth 3D world coordinates such that the mean ℓ_2 norm of valid ground-truth points equals 1. This normalization ensures a consistent geometric scale across different subjects and sequences.

Canonical coordinate maps are defined in the FLAME [30] coordinate system, where the origin is located at the center of the face. Using this canonical representation allows consistent alignment across subjects and facilitates stable learning of facial geometry.

B.2 Dataset Details

NeRSemble. For evaluation on the NeRSemble [27] dataset, we use the following five test subjects:

- 043
- 128
- 236
- 306
- 474

To evaluate different aspects of our method, we select different sequences for different tasks in order to increase evaluation diversity.

For **depth and reconstruction evaluation**, we use the following sequences:

- 043_SEN-01-cramp_small_danger
- 128_EMO-2-surprise+fear
- 236_EMO-1-shout+laugh
- 306_EXP-4-lips
- 474_EXP-3-cheeks+nose

For **tracking evaluation**, we use the following sequences:

- 043_SEN-02-same_phrase_thirty_times
- 128_EXP-9-jaw-2
- 236_EMO-4-disgust+happy
- 306_EXP-1-head
- 474_SEN-06-problems_wise_chief

Using different sequences across tasks ensures that the evaluation covers a wide range of expressions, motions, and speaking patterns.

Ava-256. For evaluation on the Ava-256 [40] dataset, we use the following subjects:

- 20210810–1306–FXN596
- 20210817–0900–NRE683
- 20210818–1332–CDR970
- 20210819–0903–DOT682
- 20210827–0906–KDA058

For all reported metrics, we restrict the evaluation to the facial region. Specifically, we use Facer [80] masks that include both the facial area and hair to define the evaluation region.

Table 8: Additional ablation study. Component analysis using depth (AbsRel) and correspondence (EPE) evaluation metrics. Best results are shown in **bold**, second-best underlined. AbsRel values are $\times 10$. We also report the average rank across all metrics (lower is better) to summarize overall performance across geometry and correspondence prediction. Our final model achieves the best average rank, indicating the most balanced performance across metrics.

Method	AbsRel (Monocular)↓	EPE↓	AbsRel (16 views)↓	Avg. Rank↓
Backbone Fixed (DA3)	0.101	6.9261	0.062	6.7
Backbone Fixed (DAViD Pretrained)	0.082	5.8219	0.040	5.3
$\lambda_C = 1$	**0.052**	4.0316	0.015	2.8
$\lambda_C = 10$	0.056	**3.1838**	0.015	**2.7**
$\lambda_D = 0$	0.064	3.2102	0.060	4.3
$\lambda_C = 0$	**0.052**	135.1671	**0.014**	3.2
Ours	0.053	3.271	0.015	**2.7**

C Additional Ablation Studies

Additional ablation study. We present additional ablation experiments in Tab. 8 to analyze the effect of backbone training and the loss weighting between depth and canonical map supervision.

First, we evaluate the effect of freezing the backbone. When the backbone is fixed and initialized with the original DA3 [33] weights, the model performs poorly on both depth and correspondence estimation, indicating that the pretrained DA3 representation is not sufficient for our task. Using a stronger initialization with a DAViD-pretrained [53] backbone improves performance across all metrics, but still remains noticeably worse than training the backbone jointly with our objectives.

Next, we analyze the impact of the canonical map loss weight λ_C. When $\lambda_C = 1$, the model struggles to learn accurate correspondences because the canonical prediction head must learn the task from scratch, while the depth head benefits from pretrained initialization. Increasing the weight to $\lambda_C = 10$ significantly improves correspondence accuracy, achieving the best EPE, but slightly degrades depth performance. Based on this trade-off, we select $\lambda_C = 5$ as a balanced setting for our final model.

We further study the effect of disabling the depth supervision by setting $\lambda_D = 0$. In this case, the model learns slightly better correspondences but the depth prediction degrades, highlighting the importance of joint geometry supervision. Conversely, removing the canonical loss ($\lambda_C = 0$) prevents the model from learning meaningful correspondences, leading to extremely large EPE values, while slightly improving the depth metrics.

Overall, our final model achieves the best balance across the evaluated metrics, obtaining the best average rank while remaining close to the top performance for each individual metric. This demonstrates that our design effectively balances geometry reconstruction and dense correspondence estimation.

D Additional Results

We present additional qualitative results demonstrating the behavior of our method on challenging in-the-wild video frames from VFHQ [71].

Additional depth and canonical predictions. Fig. 9 shows additional predictions of depth maps and canonical maps from two input views. Despite variations in pose, expression, and identity, the predicted depth and canonical coordinates remain consistent across frames. These results further illustrate the robustness of our approach when reconstructing facial geometry from unconstrained image sequences.

Canonical point cloud consistency. In Fig. 10, we visualize canonical point clouds obtained by backprojecting the predicted canonical maps into 3D space. The resulting point clouds are rendered from two viewpoints while keeping the visualization camera fixed across all samples. The results demonstrate that the canonical representation is consistent across viewpoints, facial expressions, and identities.

Additional full pipeline results. Fig. 11 presents further examples of our full pipeline, including the input images, reconstructed geometry, dense correspondences, and canonical point clouds. The predicted correspondences align well across views, highlighting the ability of our method to establish dense facial correspondences while simultaneously reconstructing temporally consistent geometry.

Failure case. We also highlight a representative failure case in Fig. 12. While the predicted correspondences remain accurate on the facial region, the method incorrectly matches the microphone visible in the scene. Since the microphone is not part of the facial surface, this leads to erroneous correspondences, illustrating a limitation of the approach when non-face objects appear in the image.

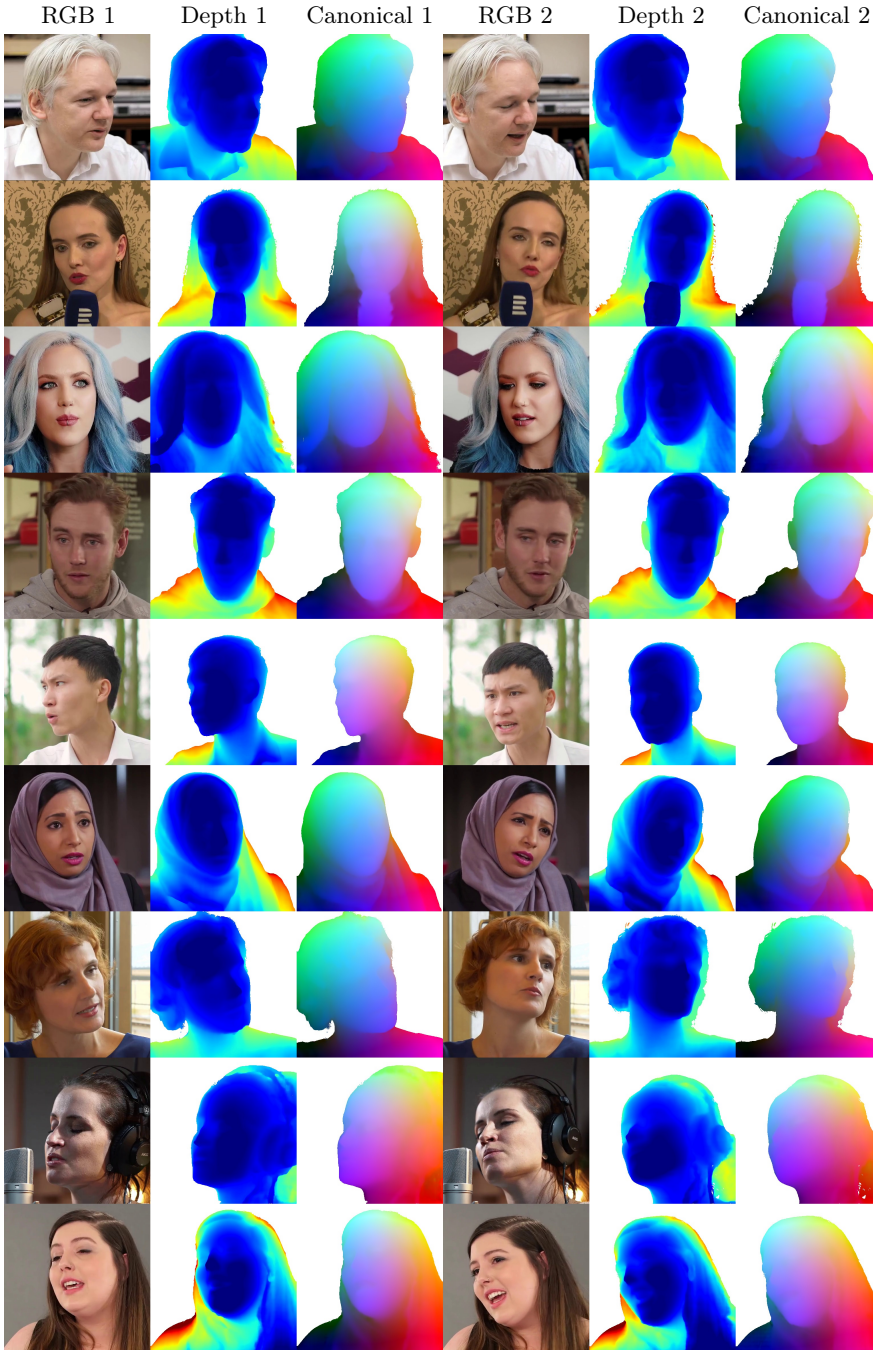

Fig. 9: Additional prediction examples on VFHQ. Given two input views, our method predicts depth maps and canonical maps for each frame. The results demonstrate consistent geometry and canonical representations across different identities and expressions.

Fig. 10: Additional canonical point cloud prediction examples on VFHQ. Given two input views, our method predicts canonical maps that are backprojected into 3D space to form canonical point clouds. We visualize the reconstructed point clouds from two viewpoints. The results show that the predicted canonical point clouds remain consistent across viewpoints, facial expressions, and identities. The visualization cameras are fixed across all samples.

Fig. 11: Additional predictions on VFHQ. Given two RGB input views, our method reconstructs 4D facial geometry and predicts dense correspondences via canonical facial coordinates. The results demonstrate consistent geometry and correspondences across different viewpoints, facial expressions, and identities.

Fig. 12: Failure case on VFHQ. Given two input RGB images, we visualize the predicted correspondences between the reconstructed point clouds. While the correspondences are largely accurate on the facial region, the method fails on the microphone, which is not part of the facial surface and leads to incorrect matches. This highlights a limitation when non-face objects are present in the scene.