Coherent 3D Portrait Video Reconstruction via Triplane Fusion

Shengze Wang[1*] Xueting Li[2] Chao Liu[2] Matthew Chan[2] Michael Stengel[2] Henry Fuchs[1]
Shalini De Mello[2†] Koki Nagano[2†]
[1] UNC Chapel Hill [2] NVIDIA † Equal Contribution
https://research.nvidia.com/labs/amri/projects/coherent3d/

Abstract

*Recent breakthroughs in single-image 3D portrait reconstruction have enabled telepresence systems to stream 3D portrait videos from a single camera in real-time, democratizing telepresence. However, per-frame 3D reconstruction exhibits temporal inconsistency and forgets the user's appearance. On the other hand, self-reenactment methods can render coherent 3D portraits by driving a 3D avatar built from a single reference image but fail to faithfully preserve the user's per-frame appearance (*e.g.*, instantaneous facial expressions and lighting). As a result, neither of these two frameworks is an ideal solution for democratized 3D telepresence. In this work, we address this dilemma and propose a novel solution that maintains both coherent identity and dynamic per-frame appearance to enable the best possible realism. To this end, we propose a new fusion-based method that takes the best of both worlds by fusing a canonical 3D prior from a reference view with dynamic appearance from per-frame input views, producing temporally stable 3D videos with faithful reconstruction of the user's per-frame appearance. Trained only using synthetic data produced by an expression-conditioned 3D GAN, our encoder-based method achieves both state-of-the-art 3D reconstruction and temporal consistency on in-studio and in-the-wild datasets.*

1. Introduction

Telepresence aims at bringing distant people face-to-face and stands out as a particularly compelling application of computer vision and graphics. Over the last decades, various successful telepresence systems [19, 21, 21, 25, 31, 33, 37, 43] have been developed. However, most employ bulky multi-view 3D scanners or depth sensors to ensure high-quality volumetric per-frame reconstruction. Unlike these classical 3D/4D reconstruction methods, recent AI-based feed-forward 3D lifting techniques [2, 52] can lift a single RGB image from an off-the-shelf webcam into a neural radiance field (NeRF) representation encoded into a set of triplanes in real time, paving the way towards making 3D telepresence accessible to anyone [48].

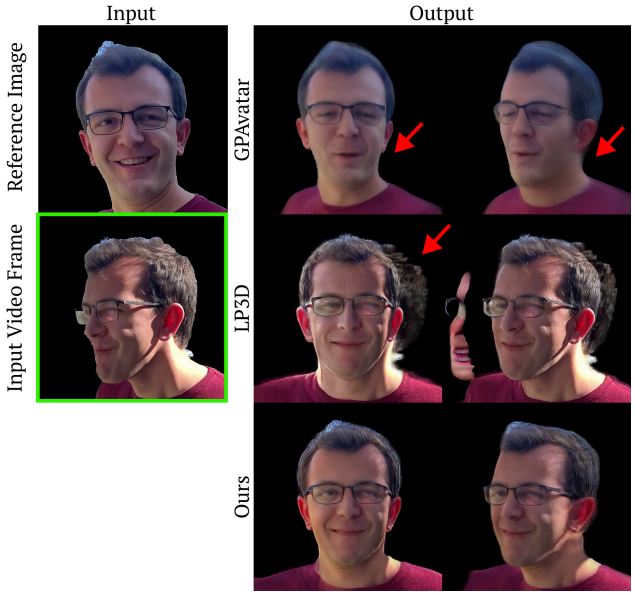

Figure 1. **Comparisons.** Given a single reference image and a single-view video frame, our method reconstructs the authentic dynamic appearance of the user (e.g., facial expressions and lighting) while producing a temporally coherent 3D video. A previous single-view 3D lifting method (LP3D) that reconstructs the avatar from the video frame on a per-frame basis suffers from distortions and temporal inconsistency. A portrait reenactment method (GPAvatar) drives the identity in the reference image using the video frame, but fails to capture accurate facial expressions (e.g., smile) and per-frame appearance (e.g., lighting). The output should be compared to the appearance of the input video frame (green box).

Currently, there are two major paradigms in democratized 3D telepresence solutions from a single-view video: (1) single-view per-frame 3D lifting methods and (2) 3D portrait reenactment, which drives an identity from a reference image using another driving frame, but none of them is an ideal solution. For (1), single-frame-based lifting techniques such as LP3D [52], have the advantage of faithfully preserving the instantaneous dynamic conditions present in an input video, *e.g.* lighting, expressions, and posture, all

*Shengze Wang was an intern at NVIDIA during the project

of which are crucial to an authentic telepresence experience. However, single-image reconstruction methods operate independently on each frame and thus have fundamental limitations in maintaining temporal consistency. This difficulty stems from the inherent ill-posed nature of single-image-based reconstruction. In order to render novel views that are significantly far from the input view, the system cannot rely on information present in the input view and hence must *hallucinate* generate plausible content, which cannot be guaranteed to be consistent across multiple temporal frames. For example, LP3D's reconstruction changes significantly depending on the user's head pose in the input frame (see second row in Fig. 1 and third column in Fig. 2). In comparison, 3D self-reenactment methods create an avatar model from one or multiple reference images and use a separate driving video to drive the facial expressions and poses of the avatar. [6, 27, 51, 59]. While these approaches allow for temporally consistent results, they often cannot faithfully reconstruct the input video's dynamic conditions such as changes in lighting. Moreover, reenactment methods often struggle to generate accurate expressions because their expression detection and control are not precise enough (see the first row in Fig. 1).

In this work, we address this dilemma of democratized 3D telepresence approaches and propose a novel solution to the problem of simultaneously maintaining temporal stability while preserving real-time dynamics of input videos for single-camera telepresence applications. Our proposed solution is a fusion-based approach that leverages the stability and accuracy of a canonical 3D prior, and also captures the diverse deviations from the prior in newer video frames (see the third row in Fig. 1).

Our model first uses LP3D [52] to construct a canonical[1] triplane prior from a (near) frontal image of the user, which can be casually captured or extracted from a video. During video reconstruction, our model lifts each input frame into a raw triplane, which is then fused with the canonical triplane (see Fig. 3). For images with oblique yaw head poses, artifacts and identity-related distortions may be present in its lifted triplane (see "LP3D" in Fig. 1 and 2). Hence we propose an undistorter module, which learns to undistort the raw instantaneous triplane to more closely match the identity of the correctly structured canonical triplane. We then propose a fuser module that combines the undistorted triplane and the canonical triplane in order to preserve identity consistency and recover occluded regions while reconstructing the dynamic lighting, nuanced expressions, and posture information in the input frame.

We summarize our contributions as follows:

- We contribute a novel triplane fusion method that combines the dynamic information from per-frame triplanes with a canonical triplane extracted from a reference image. Trained only using a synthetic multi-view video dataset, our feed-forward approach generates 3D portrait videos that demonstrate both temporal consistency and faithful reconstruction of the dynamic appearance of the user (*e.g.* lighting and expression), whereas prior solutions can only achieve one of the two properties.
- We propose a new framework to evaluate single-view 3D portrait reconstruction methods using multi-view data and gain insight into the method's reconstruction quality and robustness.
- We present evaluations on both in-studio and in-the-wild data and demonstrate that our method achieves state-of-the-art performance in terms of temporal consistency and reconstruction accuracy.

[1]We find that the 3D lifting from a near frontal reference view is reliable, hence use this as a canonical 3D prior. See Fig. 2 first column.

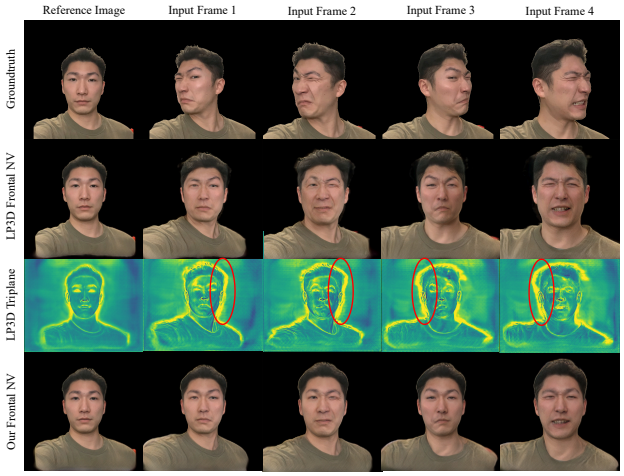

Figure 2. **View-Dependent Distortion.** "NV" refers to "Novel View Rendering". *Top*: inputs to our model and LP3D. *Second & Third Rows:* LP3D's reconstruction varies greatly under challenging viewpoints, showing a predictable pattern of artifacts including abnormally strong activations on the side being captured (red circle), as well as geometric distortion along the view direction of the camera due to depth ambiguity. We refer to this phenomenon as "View-Dependent Distortion". *Fourth row:* Our method removes such artifacts and achieves better coherence.

2. Related Work

2D portrait reenactment. Given a single or a few reference portrait images and a driving video, recent talking-head generators can reenact 2D portraits by transferring the facial expressions and poses from the driving video onto facial portraits [9, 10, 17, 46, 53, 54, 60, 62–64, 66]. However, being 2D, they cannot be rendered from novel viewpoints, which is crucial for 3D telepresence.

3D-aware portrait generation and reenactment. Some recent works use deformable volumetric implicit radiance fields [34, 38, 39] or Gaussian splatting [22] combined by 3DMMs to reconstruct a photorealistic and animatable volumetric head avatar [1, 4, 14, 41, 45, 57, 67]. However, they require extensive data captures from videos or multiview cameras and person-specific training. Others use

large-scale video datasets and learn a disentangled triplane 3D [5] for 3D facial reenactment in a feedforward fashion [6, 27, 28, 32, 51, 59, 61]. They construct a canonical 3D head from a reference image (often a neutral frame), and use facial expressions and head poses extracted from a separate driving video to animate it. As such, fine-grained facial expressions may not be captured due to errors in disentanglement. Most importantly, these reenactment methods fail to preserve the dynamic appearance of users (e.g., person-specific wrinkles or lighting) across time.

3D GAN inversion. By combining GANs [15] and neural volume rendering [34], recent breakthroughs in 3D-aware GANs [5, 8, 16, 36, 44, 47, 55, 56, 58, 65, 68] can learn to generate photorealistic 3D heads from in-the-wild 2D images. Notably, EG3D proposes an efficient and compact triplane representation [5] to generate 3D heads, and Next3D [50] extends EG3D to create 3D portrait videos controlled by 3DMM facial expression and pose parameters. Besides generating 3D heads, these models can also be used for single-view 3D reconstruction using GAN inversion [13, 24, 30, 49], manipulation of the 3D head [18, 49, 69], and 3D personalization [3, 40]. However, inverting a few seconds of video can take many minutes or hours, and the inversion quality is often unsatisfactory due to inaccurate identity and expression preservation, as evaluated by GPAvatar [6]. Therefore, recent works [2, 52] propose encoder-based solutions to lift each video frame into a triplane. However, when each frame is independently lifted into 3D, the resulting 3D head video suffers in terms of temporal consistency – a key requirement of 3D telepresence systems. To enhance single-frame 2D-to-3D encoders for human heads, *e.g.*, LP3D [52], we propose a triplane-fusion-based method, which improves their temporal stability while preserving temporal dynamics across time.

3. Method

3.1. Background: 3D Portrait from a Single Image

LP3D [52] performs photorealistic 3D portrait reconstruction by using a feedforward encoder to lift an RGB image into a triplane $\mathbf{T} \in \mathbb{R}^{3\times32\times256\times256}$, which can be volume rendered to an RGB image from any viewpoint. LP3D can run in real-time and has been developed into a complete realtime telepresence system [48]. We lift 2D faces into triplanes with a slightly modified LP3D trained on larger face crops containing shoulders and a camera estimator to recover the input image's camera parameters $M \in \mathbb{R}^{25}$. It performs better than the original version (Table 3).

3.2. Definitions

We call a current video frame an "**input frame**" into our system. Additionally, we encode a near-frontal "**reference image**" of the subject into a "**canonical triplane**", which is used to stabilize the 3D video generation process. Lastly, we refer to the viewpoint of the camera in the input frame as "**input viewpoint**".

3.3. Overview

An overview of our method is illustrated in Fig. 3. To improve temporal consistency and reconstruct occluded parts of the face, we leverage an additional near-frontal reference image, which can be a selfie image or a video frame automatically obtained from the same video using an off-the-shelf head pose estimator like DECA [11]. We lift this reference image into the canonical triplane using pre-trained LP3D. When processing the video, we first lift each input frame into a raw triplane. Then, the Triplane Undistorter (Sec. 3.5) removes view-dependent distortions and recovers the identity using the canonical triplane as a reference. This produces an undistorted triplane. Finally, to recover regions that are occluded in the input frame and further improve stability, our Triplane Fuser (Sec. 3.6) combines the undistorted triplane with the canonical triplane to generate the final coherent triplane.

3.4. Generating Synthetic Dynamic Multiview Data

We synthesize ground truth multiview training images using Next3D [50]. We first extract 2D landmarks and FLAME [26] coefficient labels for the FFHQ [20] dataset using DECA [12]. Then, during training, we sample a pair of random FLAME coefficients from our FFHQ labels and input them to Next3D alongside a single random identity code z. We notice that FFHQ contains mostly less extreme expressions. Thus, we randomly multiply the expression codes by a small scaling factor to exaggerate the expressions during training. Notice that the Next3D generator is much more expressive than FLAME due to its use of triplanes and GAN training even though it is conditioned on FLAME. Next3D then generates a pair of triplanes for $t = 0$ and $t = 1$ of the same person with 2 expressions. We render the $t = 0$ triplane into a near-frontal reference image I_{ref} (Fig. 3 orange box) and the $t = 1$ triplane into 3 images: the input frame rendered from a random viewpoint (green box), the ground truth image at a different viewpoint (red box), and a frontal-view image (red box). Only the latter two are used for supervision. Additionally, to simulate different lighting conditions, we also apply different color augmentations (brightness, contrast, saturation, and hue) to images at $t = 0$ and $t = 1$.

Shoulder augmentation. It is important that 3D portraits capture dynamic shoulder movements to convey body language and achieve eye contact in telepresence. However, Next3D does not provide control over shoulder posture. Thus, we simulate shoulder movement in the rendered images without modifying the Next3D triplanes by warping camera rays during volume rendering. Please see the supplement for more detail.

Pseudo-groundtruth triplanes. Due to our shoulder augmentation, the Next3D triplanes and their 2D renderings become inconsistent. Thus, Next3D's triplanes cannot be used

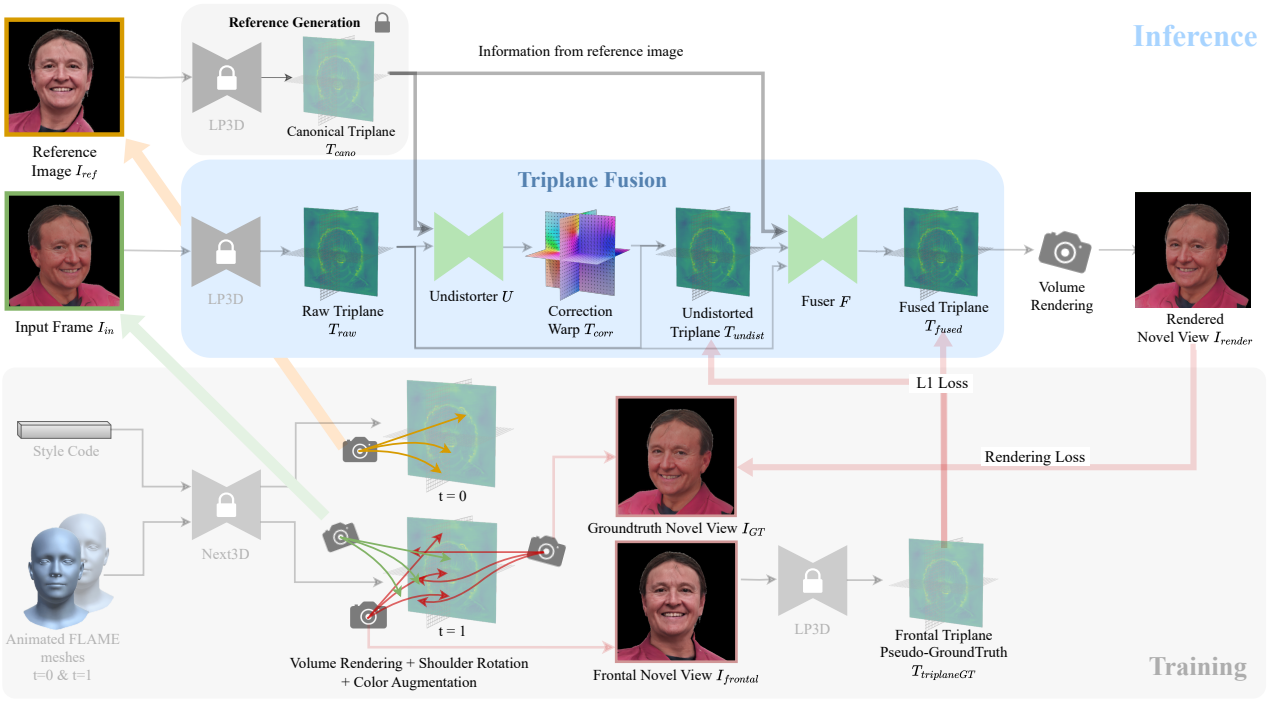

Figure 3. **Method Overview.** Given a near-frontal reference image and an input frame, we first reconstruct a canonical triplane and a raw triplane using LP3D [52] (Sec. 3.1). Next, we combine them through a Triplane Fusion module (blue box) that ensures temporal consistency while preserving dynamic information (*e.g.*, lighting and shoulder pose) (Sec. 3.5 and 3.6). Our model is trained with only synthetic video data generated by a 3D GAN [50], with carefully designed augmentations to preserve shoulder motion and lighting (Sec. 3.4).

as direct supervisory signals. To mitigate this, we leverage the fact that LP3D generates reasonably accurate triplanes from frontal view images. We use a frozen LP3D to predict pseudo-groundtruth triplanes $T_{triplaneGT}$ from the frontal novel view $I_{frontal}$ for $t = 1$ (Fig. 3 bottom right).

3.5. Removing Distortion and Preserving Identity

LP3D's reconstruction quality strongly depends on the input frame's viewpoint. When subjects are captured from the side, LP3D tends to produce errors in identity and exhibits stretching distortions (*e.g.*, in the "Input Frame 1" and "Input Frame 2" columns in Fig. 2), primarily due to the inherent ambiguity of single-image reconstruction. In contrast, near-frontal views offer more complete identity cues and suffer less from occlusion (*e.g.* the "Reference Image" column in Fig. 2), leading to more reliable reconstruction. To reduce the single-image ambiguity, we adopt a near-frontal reconstruction as the canonical triplane T_{cano}. Although incorporating more reference views could potentially improve accuracy, we find that using a single near-frontal view not only simplifies the model but also sufficiently enhances temporal coherence.

To correct the distortion in an input raw triplane T_{raw} using the canonical triplane T_{cano} as reference, we devise a Triplane Undistorter U (Fig. 3):

$$U(T_{raw}, T_{cano}) = T_{undist} \in \mathbb{R}^{3\times32\times256\times256}. \quad (1)$$

Since the distortion in each triplane is a 2D warping artifact, it can be reversed through a 2D undistortion warping of each plane. To this end, our Undistorter U adopts the SPyNet[42] architecture to predict a 2D correction warping $T_{corr} \in \mathbb{R}^{3\times2\times256\times256}$ for the three planes. Then, the Undistorter corrects the raw triplane T_{raw} by warping it based on the predicted 2D offsets $(\Delta u, \Delta v)$ in T_{corr} as:

$$T_{corr} = SPyNet(T_{raw}, T_{cano}), \quad (2)$$

$$T_{undist} = Warp(T_{raw}, T_{corr}). \quad (3)$$

It is important to note the difference between optical flow and undistortion. An ideal optical flow model would warp T_{raw} towards T_{cano} in order to make the two identical. However, this alignment leads to significant artifacts and destroys the reconstruction (Table. 3 and Fig. 5). This is because *the purpose of undistortion is not to make the two triplanes identical, but to correct T_{raw} such that it has the same identity as the canonical triplane T_{cano}, while preserving dynamic information such as expressions and lighting*. Therefore, instead of using the canonical triplane T_{cano} as the warping target, our Undistorter U merely uses T_{cano} as the identity conditioning to correct the raw triplane T_{raw} towards the ground truth triplane $T_{triplaneGT}$ (Fig. 3, lower-right), which is not available during test time. The correction warping is supervised by the consistency between the undistorted triplane T_{undist} and pseudo-groundtruth tri-

plane $T_{triplaneGT}$ (Sec. 3.4) via a triplane loss:

$$L_{undist} = L_1(T_{undist}, T_{triplaneGT}). \quad (4)$$

3.6. Reconstructing Occluded Regions through Triplane Fusion

As the user moves around in the video, different parts of their head become occluded. To recover occluded areas in the input frame and further stabilize the subject's identity across the video, our Fuser F enhances the reconstruction by utilizing the canonical triplane T_{cano}, in which the currently occluded areas are often visible. Therefore, it is important for the Fuser F to identify and recover the occluded regions while preserving information from the visible regions in T_{undist}. To accomplish this, we thus use a 5-layer ConvNet-based visibility estimator V to predict a visibility triplane as $V(T_{raw}) = T^{vis}_{raw} \in \mathbb{R}^{3\times1\times128\times128}$, *i.e.* one visibility map for each plane. T^{vis}_{raw} is undistorted alongside T_{undist} to produce T^{vis}_{undist}. We also predict a visibility triplane for the canonical triplane as $T^{vis}_{cano} = V(T_{cano})$. Finally, the Fuser F produces the fused triplane T_{fused} by combining information from the undistorted input tranplane T_{undist}, its visibility triplane T^{vis}_{undist}, the canonical triplane T_{cano}, and its visibility triplane T^{vis}_{cano}.

$$T_{fused} = F(T_{undist}, T^{vis}_{undist}, T_{cano}, T^{vis}_{cano}). \quad (5)$$

In this way, Fuser F merges new information in the input frame with occluded regions visible in the reference image. To train the visibility predictor V, we calculate the visibility loss L_{vis} as the L_1 distance between the predicted visibility triplanes versus the ground truth:

$$L_{vis} = L_1(T^{vis}_{raw}, T^{visGT}_{raw}) + L_1(T^{vis}_{cano}, T^{visGT}_{cano}). \quad (6)$$

The ground truth visibility triplanes contain 1 for pixels that are visible, and 0 otherwise. The generation of the ground truth visibility triplanes T^{visGT}_{raw} and T^{visGT}_{cano} is discussed in the supplementary.

To supervise the Fuser F, we calculate the fusion loss L_{fusion} as the L_1 loss between the fused triplane T_{fused} and the pseudo-groundtruth triplane $T_{frontalGT}$. To highlight the currently occluded region during training, we also upweight the occluded region using an occlusion mask $T_{occMask} \in \mathbb{R}^{3\times1\times256\times256}$ as:

$$T_{diff} = |T_{fused} - T_{triplaneGT}|, \quad (7)$$

$$L_{fusion} = Mean(T_{diff}) + \frac{T_{diff} \cdot T_{visGT}}{|T_{visGT}|} + \frac{T_{diff} \cdot T_{occMask}}{|T_{occMask}|}. \quad (8)$$

Please refer to the supplementary for the calculation of the occlusion mask. We use the Recurrent Video Restoration Transformer (RVRT) [29] as the backbone of our Fuser F because of its memory efficiency. We find that the final summation skip connection in RVRT prevents effective learning. This is because the original RVRT was designed to correct local blurriness and noise in a corrupted RGB video, whereas our triplane videos exhibit structural distortion on a much larger scale and the summation skip connection thus limits the model's ability to correct the general structure. We thus replace the summation with a small 5-layer ConvNet.

Lastly, note that both the Undistorter U and the Fuser F consist of 3 separate copies, one for each of the 3 planes, because we find that processing all three planes jointly leads to collapse to 2D (please see supplementary for visualization and analysis).

3.7. Training Losses

Our loss function is the summation of four loss terms that provide two types of supervision: (a) direct triplane space guidance used to supervise the undistortion process in the Undistorter U, the visibility prediction process, and the fusion process in the Fuser F; and (b) image space guidance for overall learning of high-quality image synthesis:

$$L = w_{undist}L_{undist} + w_{fusion}L_{fusion} + w_{vis}L_{vis} + w_{render}L_{render}, \quad (9)$$

where w_{undist}, w_{viz} , w_{fusion}, and w_{render} are scalar weights for the different loss terms. L_{render} is calculated as the perceptual loss L_{LPIPS} between the ground truth novel view I_{GT} and the rendered novel view I_{render} as:

$$L_{render} = L_{LPIPS}(I_{GT}, I_{render}). \quad (10)$$

4. Evaluation

Multi-view evaluation of single-view reconstruction. Due to the lack of 3D ground truth for in-the-wild images, prior methods are often evaluated on the input image reconstruction task, which compares the rendered image against the input image using metrics like PSNR, whereas the novel view synthesis task often relies on visual assessments. However, this practice can lead to ambiguities and inaccurate conclusions. For example, if a 3D reconstruction is only rendered and evaluated from the input viewpoint, then that reconstruction can overfit to the input view to achieve high scores even if it is inaccurate when rendered from other viewpoints (as proven quantitatively in Fig. 11 of the supplementary). Moreover, single-view reconstruction methods can be heavily affected by the choice of input viewpoints (Fig. 2). Therefore, both the input viewpoint and the novel viewpoint are crucial variables to consider when concluding the performance of a method. We thus propose a new evaluation framework that evaluates a model across every input-novel viewpoint combination. In this way, a method can only achieve high numerical performance when it consistently generates high-quality reconstructions regardless of the choice of input or novel viewpoint.

4.1. Evaluation Dataset

We quantitatively evaluate various methods on the NeRSemble [23] dataset, which is a multiview portrait

video dataset that allows us to evaluate the methods using different input-novel viewpoint combinations. It is recorded with 16 calibrated time-synchronized cameras in a controlled studio environment. The images are captured at 7.1 MP resolution and 73 frames per second. There are 10 sequences in the test set, each capturing a different individual performing different expressions. One of the 10 test sequences involves severe facial occlusion from hair that causes most of the methods' face trackers to fail for significant portions of the recording for many of the viewpoints. We thus leave out that sequence because the results would not be a reliable assessment of quality. We also use 8 roughly evenly separated cameras out of all 16 cameras during evaluation.

4.2. Metrics

Synthesis Quality. Given N views in the dataset, we evaluate a method's average performance across every input-novel viewpoint combination. More specifically, for each frame, each of the N cameras is used as the input viewpoint, producing N reconstructions in total. Then, each of the N reconstructions is rendered and evaluated on the N viewpoints, resulting in an $N \times N$ score matrix. As mentioned before, we use $N = 8$ of the camera views from the NeRSemnble[23] dataset as input and novel views. Thus, for a test sequence of T frames, we can calculate a spatial-temporal score matrices $\mathbf{S}^{T\times 8\times 8}$ for each of the metrics (as visualized in Fig. 11 in the supplementary):

$$\mathbf{S}_{t,i,j} = Metric(\mathbf{I}^{t,i,j}_{render}, \mathbf{I}^{t,j}_{GT}), 1 \leq i,j \leq N, 1 \leq t \leq T, \quad (11)$$

$$s = Mean(\{\mathbf{S}_{t,i,j}\}), \quad (12)$$

where $Metric(\cdot)$ can be LPIPS, PSNR, *etc.*. $\mathbf{I}^{t,i,j}_{render}$ is the image rendered using camera i as the input frame and camera j as the output rendering view at frame t. $\mathbf{I}^{t,j}_{GT}$ is the ground truth frame captured by camera j at frame t. The Synthesis Quality s is thus the average over all entries in $\mathbf{S}$. For a dataset with multiple sequences, such as NeRSemble, the final Synthesis Quality is the average across all sequences.

Novel View Synthesis (NVS) Quality. Novel View Synthesis Quality s_{NV} is the average over all entries where the input view i is different from the novel view j, essentially removing the diagonal entries from the score matrix:

$$s_{NV} = Mean(\{\mathbf{S}_{t,i,j} | i \neq j, 1 \leq t \leq T\}). \quad (13)$$

Identity Accuracy (ID) We measure the ArcFace [7] cosine distance between the ground truth image I_{GT} and rendered image I_{render} both from the frontal camera.

Expression Accuracy ($Expr$) We use NVIDIA Maxine AR SDK [35] to measure the L_2 distance between expression coefficients e_{GT} of the ground truth image and e_{render} of the rendered image both from the frontal camera.

Dynamic Appearance While we believe that it is important to measure the accuracy of dynamic appearance such as lighting and shoulder poses, there is no existing multi-view in-the-wild portrait video dataset to support such evaluation. We thus qualitatively evaluate various methods on challenging in-the-wild portrait videos. Please see the supplementary material for image examples and video results.

4.3. Comparisons

Baselines. We compare recent methods from 3 categories: (i) *Reconstruction:* We evaluate LP3D [52] using the above protocol. LP3D generates N 3D portraits using each of the N viewpoints as input, and the N 3D portraits are then evaluated on the N ground truth viewpoints. (ii) *Reenactment:* We evaluate Li *et al*. [28] and GPAvatar [6] in the self-reenactment setting. Li *et al*. [28] reconstruct 3D portraits as triplanes from reference images without test-time optimization, and they drive the 3D portraits via dynamic frontal renderings of 3DMMs. The authors of Li *et al*. [28] evaluated their approach using the input views as the only novel view, instead of all N views. GPAvatar[6] reconstructs 3D portraits by leveraging multiple source images and driving them through a FLAME [26] mesh model. We use the first frame of the frontal camera in each NeRSemble test sequence as the reference image to generate the 3D portrait, and drive it using videos from each of the N viewpoints. We evaluate GPAvatar using the same evaluation protocol as our method and LP3D. (iii) *Inversion:* We evaluate VIVE3D [13], which is a state-of-the-art 3D GAN inversion method for videos, and it can also perform semantic video editing. We first provide VIVE3D 3 video frames from the input viewpoint, which it requires for personalization. Then, we invert the input video frames into 3D portraits and render the 3D portraits from all N viewpoints. Each of the above method uses a different cropping of the face. We standardize the evaluation by re-cropping all methods to our cropping protocol, which is the largest of all. Please see the supplement for an additional table, where we evaluate the methods using different croppings around the face and arrive at conclusions consistent with Table 1.

Quantitative results. We evaluate various methods using different input-novel viewpoint combinations, providing a robust multi-view assessment of a model's performance. Table. 1 shows that our model achieves state-of-the-art performance across all metrics versus recent works. Notably, while LP3D overfits to the input viewpoint (as shown by the difference of 0.806dB PSNR between Synthesis Quality and NVS Quality), our method generalizes better to novel viewpoints (as shown by the difference of 0.330dB). Moreover, our method best preserves subject identity and expression (Table. 1, Fig. 2, and 4). On the other hand, the reenactment methods struggle to capture authentic expressions because of the use of morphable face models, which have limited expressiveness and accuracy. Moreover, they cannot reconstruct dynamic appearance (*e.g.* the stuck-out tongue in the second and third rows of Fig. 4) because they solely rely on information present in the reference image(s) and

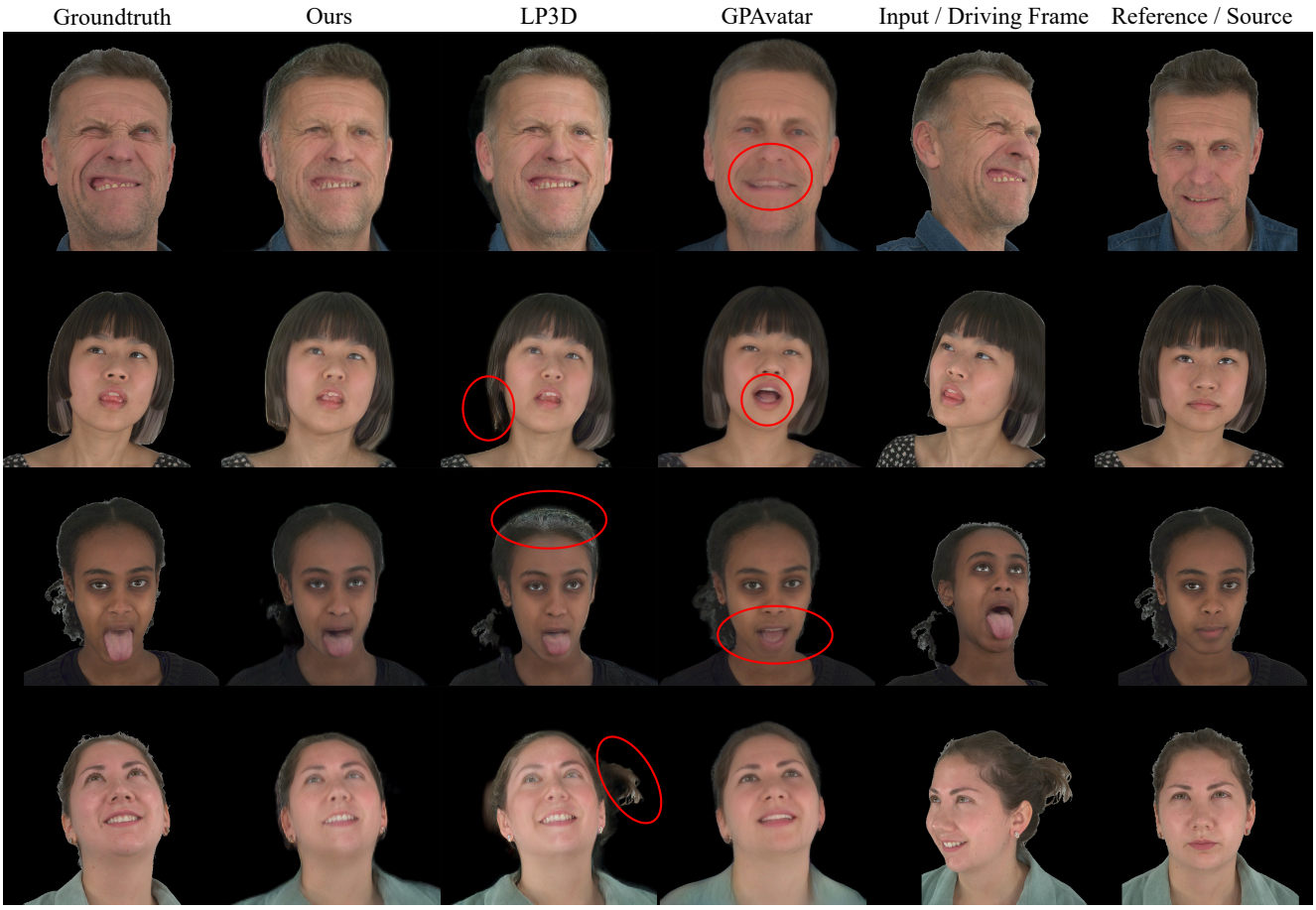

Figure 4. **Visual comparisons with baseline methods.** Our method strikes a balance between coherent reconstruction and faithfully preserving dynamic conditions like expressions. LP3D (third column) exhibits inconsistency in identities, hairstyles, and artifacts (red circles). GPAvatar (fourth column) fails to capture challenging expressions (first row), new information not present in the reference image, (the stuck-out tongue in second and third rows), and the identity of the person (last row).

do not incorporate new per-frame information. On the other hand, our method achieves faithful dynamic appearance and coherent reconstruction at the same time.

Method	Type	Expr↓	ID↓	Synthesis Quality		NVS Quality	
				PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
Li *et al*. [28]	reenact	0.266	0.241	18.573	0.255	18.202	0.262
GPAvatar[6]	reenact	0.204	0.207	21.949	0.233	21.949	0.2334
VIVE3D[13]	invert	0.290	0.395	18.577	0.259	18.145	0.271
LP3D[52]	recon	0.168	0.215	22.331	0.223	21.525	0.237
Ours	recon	**0.158**	**0.187**	**22.770**	**0.219**	**22.440**	**0.224**

Table 1. **Comparisons on Nersemble [23].** Our evaluation protocol (Sec. 4.2) utilizes multiview groundtuth for thorough analysis. Our method achieves state-of-the-art performance across all metrics, maintains temporal coherence, and preserves accurate expressions.

4.4. Ablations

We evaluate the effectiveness of our proposed Triplane Undistorter and Fuser modules on NeRSemble. In addition to the Synthesis Quality in terms of PSNR and LPIPS, we develop two new metrics to separately measure the robustness to variations in (1) rendering viewpoints and (2) input viewpoints:

(1) **Novel View Variation (NVV).** We evaluate how much a method's reconstruction quality varies across different novel views. We quantify this as the standard deviation of performance across the N novel views when using the same input view, *i.e*. the average standard deviation of each horizontal row of the score matrix S:

$$\text{NVV} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\text{stddev}\Big(\{S_{t,i,j}|1 \leq j \leq N\}\Big). \quad (14)$$

(2) **Input View Variation (IVV).** We measure how much a method's reconstruction quality varies when using the N different input viewpoints (Sec. 4.2 second column from the

Method	U	F	Synthesis Quality		IVV↓		NVV↓	
			PSNR↑	LPIPS↓	(PSNR)	(LPIPS)	(PSNR)	(LPIPS)
LP3D[52]	✗	✗	22.331	0.223	1.025	0.015	2.200	0.053
Ours	✓	✗	22.196	0.221	0.907	0.009	1.699	0.038
	✗	✓	22.265	0.223	0.559	0.006	**1.315**	**0.029**
	✓	✓	**22.770**	**0.219**	**0.245**	**0.005**	1.383	0.037

Table 2. **Undistorter and Fuser Ablations.** The Undistorter-only (row 2) and Fuser-only (row 3) variants both lead to improvements in viewpoint robustness metrics (IVV and NVV), but not in Synthesis Quality. The best performance is achieved by using both the Undistorter and the Fuser at the same time.

right). We quantify this variation as the average standard deviation of performance on the same novel view when using different input views, *i.e*. the average standard deviation of each vertical column of the score matrix S:

$$\text{IVV} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{j=1}^{N}\text{stddev}\Big(\{S_{t,i,j}|1 \leq i \leq N\}\Big). \quad (15)$$

NVV and IVV are measured for both PSNR and LPIPS, and they are better if lower.

As shown in Table. 2, the Undistorter module consistently improves the NVV and IVV metrics versus LP3D, indicating better robustness to different input viewpoints and more consistent rendering quality across novel rendered views. However, when only the Undistorter is added (Tab. 2 second row) the PSNR is slightly reduced. This is likely because this model does not leverage the reference image to recover occluded areas while also overfitting to the input view. Similarly, the Fuser-only variant (Tab. 2 third row) achieves better NVV and IVV scores, but also scores lower in terms of PSNR. A likely cause is that, without the Undistorter, the Fuser needs to merge highly misaligned triplanes, where the person looks drastically different in T_{raw} and T_{cano}. This misalignment possibly induces more blurriness and alignment artifacts that lower the PSNR. The best performance is achieved by our full model, which includes both the Undistorter and the Fuser and thus achieves better PSNR, LPIPS, IVV and NVV scores. This means that the Undistorter and the Fuser complement each other: The Undistorter corrects the distortion in the raw triplane and thus reduces the challenges in fusing misaligned triplanes, and the Fuser recovers the occluded areas in the raw triplane.

Shoulder Augmentation. without shoulder augmentation, the model fails to correctly reconstruct varying shoulder poses, resulting in worse performance (Fig. 5 and Table. 3).

Optical Flow. As mentioned in Sec. 3.5, replacing the Undistorter with optical flow worsens the reconstruction (Table. 3). As shown by Fig. 5, this variant fails to correctly reconstruct the shoulder and the mouth accurately.

GAN Training. We retrained our model with an additional adversarial loss incorporated into Eqn. 10. As expected, the GAN training (Table. 3) improves the sharpness of images as seen from the LPIPS error reduction from 0.219 to 0.207. Additional results are included in the supplementary.

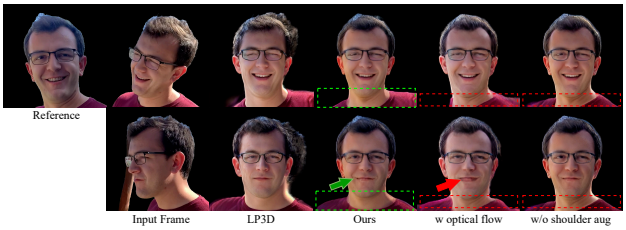

Figure 5. **Visual ablation.** Our method with optical flow and w/o shoulder augmentation on two different input frames (top and bottom rows).

Method	LPIPS↓	PSNR↑	IVV (PSNR) ↓	ID↓	Expr↓
LP3D (orig.)	0.334	18.721	2.130	0.247	0.451
LP3D (ours)	0.223	22.331	1.025	0.168	0.215
w optical flow	0.227	22.085	1.175	0.178	0.335
w/o shoulder aug.	0.218	22.342	0.829	0.153	0.244
Ours	0.219	**22.770**	**0.245**	0.158	**0.187**
Ours + GAN	**0.207**	22.074	0.314	**0.152**	0.198

Table 3. **Other Ablations.** We compare our implementation of LP3D to the original one [52] (row 1&2); the effect of replacing our Undistorter with optical flow warping (row 4); the effect of removing shoulder augmentation during training (row 4); and the effect of adding an adversarial GAN loss (row 6) alongside Eqn. 10.

Qualitative results. In-the-wild experiments show that our method achieves better temporal consistency than LP3D [52] and more accurately captures dynamic information like expressions and lighting changes than GPAvatar [6]. We refer the readers to our supplementary materials for visual examples of these results.

5. Discussion

Conclusion. Recognizing the individual limitations of per-frame single-view reconstruction and 3D reenactment methods, we presented the first single-view 3D lifting method to reconstruct a 3D photorealistic avatar with faithful dynamic appearance as well as temporal consistency, marrying the best of both worlds. We believe our method paves the way forward for creating a high-quality telepresence system accessible to consumers.

Limitations and future work. We use a single reference image, but incorporating multiple ones with different expressions and head poses could lead to further improvements. With our method, fusing an extreme side view with a very different expression to the reference view may result in blurry reconstruction due to ambiguity in triplane alignment. Finally, due to the additional components, our current run-time performance is slower than real time, which could be improved in future work.

References

[1] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[2] Ananta R. Bhattarai, Matthias Nießner, and Artem Sevastopolsky. Triplanenet: An encoder for eg3d inversion. 2024. 1, 3

[3] Marcel C. Buehler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, and Abhimitra Meka. Preface: A data-driven volumetric prior for few-shot ultra high-resolution face synthesis. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

[4] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (SIGGRAPH)*, 2022. 2

[5] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[6] 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 *International Conference on Learning Representations (ICLR)*, 2024. 2, 3, 6, 7, 8

[7] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[8] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[9] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[10] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. *arXiv preprint arXiv:2207.07621*, 2022. 2

[11] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. 2021. 3

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

[13] Anna Frühstück, Nikolaos Sarafianos, Yuanlu Xu, Peter Wonka, and Tony Tung. VIVE3D: Viewpoint-independent video editing using 3D-Aware GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 6, 7

[14] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 3

[16] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. StyleNeRF: A style-based 3D-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 3

[17] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. 2022. 2

[18] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[19] Andrew Jones, Magnus Lang, Graham Fyffe, Xueming Yu, Jay Busch, Ian McDowall, Mark Bolas, and Paul Debevec. Achieving eye contact in a one-to-many 3d video teleconferencing system. *ACM Transactions on Graphics (SIGGRAPH)*, 2009. 1

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[21] Peter Kauff and Oliver Schreer. An immersive 3d video-conferencing system using shared virtual team user environments. In *Proceedings of the 4th International Conference on Collaborative Virtual Environments*, 2002. 1

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (SIGGRAPH)*, 2023. 2

[23] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 2023. 5, 6, 7

[24] Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 3d gan inversion with pose optimization. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023. 3

[25] Jason Lawrence, Dan B Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G. Desloge, Tommy Fortes, Eric M. Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, Claude Knaus, Brian Kuschak, Ricardo Martin-Brualla, Harris Nover, Andrew Ian Russell, Steven M. Seitz, and Kevin Tong. Project starline: A high-fidelity telepresence system. *ACM Transactions on Graphics (SIGGRAPH ASIA)*, 2021. 1

[26] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. 2017. 3, 6

[27] Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang,

Liefeng Bo, and Xuelong Li. One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[28] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot neural head avatar. 2023. 3, 6, 7

[29] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. 2022. 5

[30] C.Z. Lin, D.B. Lindell, E.R. Chan, and G. Wetzstein. 3d gan inversion for controllable portrait image animation. In *ECCV Workshop on Learning to Generate 3D Shapes and Scenes*, 2022. 3

[31] S. Ma, T. Simon, J. Saragih, D. Wang, Y. Li, F. La Torre, and Y. Sheikh. Pixel codec avatars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[32] Zhiyuan Ma, Xiangyu Zhu, Guojun Qi, Zhen Lei, and Lei Zhang. Otavatar: One-shot talking face avatar with controllable tri-plane rendering. 2023. 3

[33] Andrew Maimone, Jonathan Bidwell, Kun Peng, and Henry Fuchs. Enhanced personal autostereoscopic telepresence system using commodity depth cameras. *Computers & Graphics*, 2012. 1

[34] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[35] NVIDIA. Nvidia maxine ar sdk. https://github.com/NVIDIA/MAXINE-AR-SDK, 2024. 6

[36] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[37] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. Holoportation: Virtual 3d teleportation in real-time. In *UIST 2016*, 2016. 1

[38] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. 2021. 2

[39] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[40] Luchao Qi, Jiaye Wu, Annie N. Wang, Shengze Wang, and Roni Sengupta. My3dgen: A scalable personalized 3d generative model, 2023. 3

[41] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. 2024. 2

[42] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 4

[43] Ramesh Raskar, Greg Welch, Matt Cutts, Adam Lake, Lev Stesin, and Henry Fuchs. The office of the future: a unified approach to image-based modeling and spatially immersive displays. In *ACM Transactions on Graphics (SIGGRAPH)*, 1998. 1

[44] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[45] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[46] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[47] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[48] Michael Stengel, Koki Nagano, Chao Liu, Matthew Chan, Alex Trevithick, Shalini De Mello, Jonghyun Kim, and David Luebke. AI-Mediated 3D Video Conferencing. In *ACM SIGGRAPH 2023 Emerging Technologies*, 2023. 1, 3

[49] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Transactions on Graphics (SIGGRAPH ASIA)*, 2022. 3

[50] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 4

[51] Phong Tran, Egor Zakharov, Long-Nhat Ho, Anh Tuan Tran, Liwen Hu, and Hao Li. Voodoo 3d: Volumetric portrait disentanglement for one-shot 3d head reenactment. 2024. 2, 3

[52] Alex Trevithick, Matthew Chan, Michael Stengel, Eric R. Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. In *ACM Transactions on Graphics (SIGGRAPH)*, 2023. 1, 2, 3, 4, 6, 7, 8

[53] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[54] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[55] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. *arXiv preprint arXiv:2206.07255*, 2022. 3

[56] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[57] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[58] Zhongcong Xu, Jianfeng Zhang, Junhao Liew, Wenqing Zhang, Song Bai, Jiashi Feng, and Mike Zheng Shou. Pv3d: A 3d generative model for portrait video generation. In *The Tenth International Conference on Learning Representations*, 2023. 3

[59] Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiangwei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, Chen Zhang, Xiang Yin, Zejun Ma, and Zhou Zhao. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 3

[60] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. 2022. 2

[61] Wangbo Yu, Yanbo Fan, Yong Zhang, Xuan Wang, Fei Yin, Yunpeng Bai, Yan-Pei Cao, Ying Shan, Yang Wu, Zhongqian Sun, and Baoyuan Wu. Nofa: Nerf-based one-shot facial avatar reconstruction. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 3

[62] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[63] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision (ECCV)*, 2020.

[64] Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, Hsiang-Tao Wu, Dong Chen, Qifeng Chen, Yong Wang, and Fang Wen. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation, 2023. 2

[65] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Pan Pan, and Yi Yang. Multi-view consistent generative adversarial networks for 3d-aware image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[66] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3657–3666, 2022. 2

[67] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[68] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis. *arXiv preprint arXiv:2110.09788*, 2021. 3

[69] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *European Conference on Computer Vision (ECCV)*, 2022. 3