

PAV: Personalized Head Avatar from Unstructured Video Collection

Akin Caliskan¹, Berkay Kicanaoglu¹, and Hyeongwoo Kim²

¹ Flawless AI

² Imperial College London

Abstract. We propose PAV, Personalized Head Avatar for the synthesis of human faces under arbitrary viewpoints and facial expressions. PAV introduces a method that learns a dynamic deformable neural radiance field (NeRF), in particular from a collection of monocular talking face videos of the same character under various appearance and shape changes. Unlike existing head NeRF methods that are limited to modeling such input videos on a per-appearance basis, our method allows for learning multi-appearance NeRFs, introducing appearance embedding for each input video via learnable latent neural features attached to the underlying geometry. Furthermore, the proposed appearance-conditioned density formulation facilitates the shape variation of the character, such as facial hair and soft tissues, in the radiance field prediction. To the best of our knowledge, our approach is the first dynamic deformable NeRF framework to model appearance and shape variations in a single unified network for multi-appearances of the same subject. We demonstrate experimentally that PAV outperforms the baseline method in terms of visual rendering quality in our quantitative and qualitative studies on various subjects.

1 Introduction

Digital human avatars enabling the editing of facial expressions and head motion have broad applications across telepresence, animation, and digital content creation. The demands for personalized human avatars such as digital doubles have also increased over the past few years not only in the film industry but in the research community. To enable the widespread adaptation of these neural avatars, they should be easy to generate and animate under novel poses and facial expressions. A vast majority of existing methods aim to learn neural avatars from monocular videos to make it more accessible and easy to use [1, 10, 12–14, 16, 34, 42, 44]. Early approaches [12, 13, 16, 34] mainly rely on 3D morphable models (3DMM) to guide the modeling process and to provide an intuitive control mechanism in novel view synthesis. Despite their success, they do not model the 3D face geometry. To address this issue, neural radiance fields (NeRFs) [23] have demonstrated volumetric 3D modeling of face geometry and appearance [1, 10, 14, 42, 44] under novel poses and expressions. While this allows

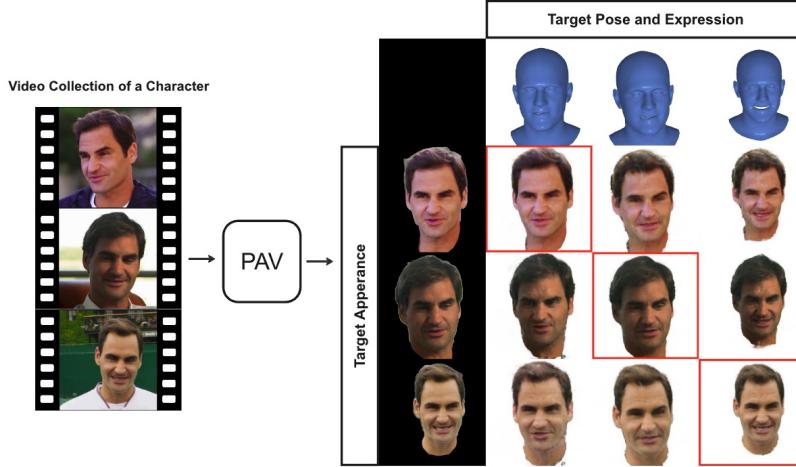


Fig. 1: We present an approach for learning personalized head avatars from an unstructured video collection of the same subject. The videos are taken in different environments, with diverse appearances and possible shape changes of a character, as shown on the left. Our method leverages all of the input clips into a single unified dynamic deformable NeRF framework and offers high-quality video replays in novel head poses and facial expressions. On the right, we show some rendering results of each input appearance of the same subject aligned in head pose and facial expression that are not seen during training. Notably, our method can synthesize identity-specific details while showing coherent expressions across different appearances.

photo-realistic image quality and animation, NeRF-based approaches require per-appearance optimization, maintaining the consistent appearance of the subject, which is constrained to the pre-recording of characters in controlled settings. However, this limits the potential for practical application scenarios where we want to generate a 3D head avatar from a collection of prerecorded video clips of a subject. In this work, we present the Personalized Head Avatar method transforming an unstructured personal video collection, containing images spanning multiple years with different appearances, head poses, and expressions, into a 3D representation of the subject. To create an avatar, we only require an unstructured collection of video clips of the person in arbitrary poses, expressions, and appearances. Our novel framework enables creation of avatars from sparse observations, while yielding high reconstruction quality in novel combinations of pose and appearance, and opens up new applications.

Learning dynamic deformable head NeRF from unstructured video clips is a challenging problem. First, video collections can contain frames of the same identity at different appearances, poses, and expressions. Second, we only have limited pose and expression variation for each appearance, so it is unlikely that all poses and expressions of the face would be well-observed for any given appearance. Given the data, existing NeRF approaches' appearance-specific nature

largely limits their potential for practical application scenarios. They require an independent model that is trained from scratch for every appearance from each monocular video. To address these challenges, we model shared deformation field and canonical space neural volumetric representation for all appearances with appearance-dependent latent features. Our key insight is that although observed faces have different appearances across video clips, they should all be explained by a common deformation model with small differences since they all originate from the same person. Furthermore, we assert that every video clip specific to an appearance contains both general and unique information. A single unified network could collaboratively learn the generic information from various appearances, thus this benefits rendering performance as well. The single network is designed to learn a deformation-aware dynamic head NeRF while disentangling appearance-related details from those that are not appearance-specific. For appearance conditioning, we propose employing *learnable appearance embedding* attached to vertices of the tracked mesh and passing them to the dynamic head NeRF network. By adopting this process, PAV is able to learn discriminative latent neural features that are specific for each appearance, resulting in better rendering of appearance-specific details for arbitrary poses and expressions. Unlike existing NeRF methods [1, 10, 14, 42, 44], PAV simultaneously trains on all appearances of the same subject, eliminating the need for separate training sessions for each appearance and reducing the overall training duration. Learning of personalized head avatars in a single unified network from an unstructured collection of videos has not been addressed before and is a non-trivial extension of existing monocular head avatar generation methods since they can not handle multiple appearances in a single network. PAV makes this possible through a carefully designed appearance conditioning representation, geometry-attached trainable latent neural texture to learn appearance-specific features, appearance conditioned implicit functions to predict variations in color and density. We conducted a variety of experiments to show these improvements over the baseline approach. Since there is no such public dataset to evaluate our method, we curate our own in-the-wild video collections for multiple subjects. Overall, we present Personalized Head Avatar with the following contributions:

- We present, PAV, a framework for learning controllable head avatars from unconstrained short video collections where the same person shows different appearances and facial expressions. To the best of our knowledge, this is the first method learning appearance controllable head avatars from in-the-wild monocular video collections of the same subject.
- The proposed method can synthesise images of head avatar under target appearance, head pose and facial expressions.
- PAV learns modeling local changes of geometry and color across different appearances of the same identity.

2 Related Work

Monocular Neural Head Avatars. More recent methods learn neural representations of the person from monocular videos and use neural renderers [31, 32] to directly generate photo-realistic images in the target head pose and facial expression [1–3, 5, 10, 11, 14, 16, 21, 29, 33, 40, 42, 44]. These methods are generally classified into image-based or 3D-based neural rendering methods [31, 32]. The image-based models synthesise the face of a subject without relying on any representation of 3D space. Existing image-based approaches either employ learned warping fields [2, 29] to deform an input image or encoder-decoder architecture where decoder synthesize the output image [21, 25, 35, 39]. Even though these methods produce high-quality results, they suffer from artifacts for strong head pose and expression changes due to the lack of 3D geometry. The 3D-based neural rendering methods, on the other hand, rely on 3D Morphable Models (3DMM) [9]. These methods are generally classified into explicit and implicit models. Explicit models utilize 2D neural rendering methods [31, 32] and rely on neural networks which can be conditioned on coarse RGB renderings based on a linear texture model [16], uv-maps [4] or latent feature maps [33]. Although these techniques produce high-quality renderings, they are confined to operating within a 2D image space and are limited to craniofacial structures. To tackle this, implicit neural representations for modeling 3D faces have recently become a major focus of attention. In particular, NeRFs [23] shows photorealistic face image synthesis in novel poses and expressions [1, 10, 14, 42, 44]. Implicit models represent the geometry using implicit surface functions or volumetric representations and they are optimized via volumetric rendering [23]. Early approaches [23] and follow-up works synthesize novel views of complex static scenes using differentiable volumetric rendering. After introduction of NeRF for static scenes, a natural research direction was to generalize it to dynamic and time-varying ones [26, 28]. The reconstruction problem is divided into two different spaces, the deformed space and the canonical space. For human body, a series of approaches have been proposed to target this challenge [20, 27]. In particular, these methods are differentiated from each other based on deformation field modeling. Gafni et al. [10] introduced implicit deformation modeling conditioned on facial expression and head pose. In contrast to modeling the deformation field implicitly, IMAvatar [42], RigNerf [1] and Insta [44] propose explicit modeling of the deformation field to improve control and rendering quality. Our method falls into this paradigm as we build our system on top of dynamic deformable head NeRF. In particular, we take inspiration from Insta [44] for designing our framework. However, existing deformable dynamic head NeRF methods including Insta [44] require per-identity optimization, maintaining the consistent appearance of the same subject. To the best of our knowledge, PAV is the first method to learn a single unified dynamic deformable head NeRF from monocular video of multiple appearances of the same subject.

Geometry-anchored Features for NeRFs. PAV introduces appearance-conditioned head NeRF by using geometry-attached features. By using these features, exist-

ing works show improved rendering quality for static objects [38] and human bodies [3, 20]. However, none of the existing methods [3, 20] has yet addressed multi-appearance learning for dynamic neural radiance fields.

NeRF from Unstructured Data Collections. Learning neural radiance field from unstructured data collections has recently gained much attention. Existing works have enabled photorealistic view synthesis of challenging static scenes, including tourist sites [22] and city-scale scenes [30], and static objects [18] from internet photo collections. For human subjects, Weng et al. [37] introduce free-viewpoint human body rendering from an unstructured photo collection. Despite the photorealistic rendering quality of PersonNerf [37], they do not model animatable faces due to the static training images. In our work, we learn animatable 3D head avatars from video collections.

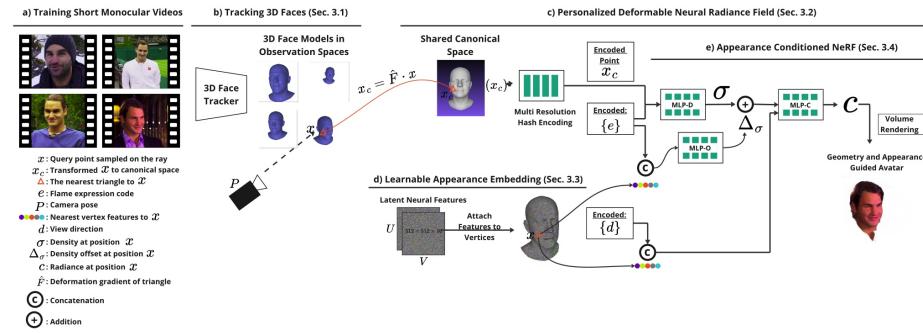


Fig. 2: Overview of the proposed approach, PAV. (a) shows some training videos in various appearances of the same subject. (b) We estimate the 3D head geometry of the person using FLAME head model. (c) shows the proposed personalized deformable neural radiance field from video collection (d) We learn appearance-specific neural features and attach them to vertices. (d) This feature is passed to fully fused multi-layer perceptrons with additional conditioning on the facial expressions e and the encoded view direction d .

3 Method

Our goal is to learn a personalized head avatar, PAV, from a monocular video collection of a subject under different appearances, poses, and facial expressions. To create the avatar, we use a small collection $\mathbf{V} = \{V_i\}_{i=1}^N$ with N videos. We parameterize PAV using FLAME [19] face model to control head poses, facial expressions and use *neural texture* to model appearance. During inference, we only need the target FLAME parameters and *neural texture*. Learning PAV from a collection of monocular videos requires capturing the geometry and appearance of the dynamic head while also disentangling different appearances of the same

identity. In order to tackle this problem, we first capture the dynamic 3D geometry by employing a parametric head model [19] (Sec. 3.1). We use the coarse geometry to model geometry-guided deformable neural radiance field (Sec. 3.2) and vertex-attached neural features to condition appearance for rendering (Sec. 3.3). We also propose estimating appearance-conditioned neural radiance field, aimed at accurately capturing distinct color and geometry with high fidelity. (Sec. 3.4). An overview of our method can be seen in Fig. 2. In the following, we describe each of these modules in greater detail.

3.1 Head Geometry Estimation

Given the training frames $\mathbf{I}=\{I_i\}$, from all appearances, we first estimate the optimized camera parameters $\mathbf{K} \in \mathbb{R}^{3 \times 3}$, tracked FLAME [19] meshes $\mathbf{M}=\{M_i\}$ with corresponding facial expression coefficients $\mathbf{E}=\{E_i\}$ and poses $\mathbf{P}=\{P_i\}$. We follow an optimization-based method, that minimizes an objective function with photo-consistency and landmark terms [43]. Different from the original work [43] which is designed for a single appearance, in PAV, we compute the average shape parameters from all appearances of the same subject to represent identity shape during optimization.

3.2 Personalized Deformable NeRF

Our goal is to create a head avatar that can be learned from frames in video collection of various appearance of the same subject. We parameterize PAV using FLAME model and learned appearance conditioning. Specifically, PAV takes target *head pose* θ , *facial expression* e and *latent neural features (LNF)* $Z_j \in \mathbb{R}^{512 \times 512 \times 16}$ in the texture atlas space (UV space) for an appearance j as inputs and synthesizes the target image I^{θ, e, Z_j} as:

$$I^{\theta, e, Z_j} = \text{PAV}(\theta, e, Z_j, K), \quad (1)$$

where K corresponds to the intrinsic parameters of the target camera viewpoint. For this purpose, we are using a geometry-guided dynamic deformable neural radiance field. However, deformable head NeRF models cannot disentangle appearance from the neural radiance field (See Sec. 2). To tackle this problem, we employ shared canonical space for all appearances where a shared neural radiance field is constructed. With this shared radiance field at canonical space, to render the target head pose and facial expression using volumetric rendering, we canonicalize the sample points x on a ray from the observation space to query points x_c in the shared canonical space radiance field by using deformation gradient \hat{F} : $x_c = \hat{F} \cdot x$ where $\hat{F} \in \mathbb{R}^{4 \times 4}$ and $\{x, x_c\} \in \mathbb{R}^4$. We employ the nearest triangle search for a given point to compute the deformation gradient. Specifically, the deformation gradient is computed between the deformed triangle $\hat{T}_{def} \in M_i$ in the observed mesh and the canonical triangle $\hat{T}_{canon} \in M^{canon}$ in the canonical mesh. To increase the speed of nearest triangle search, we employ *bounding volume hierarchy (BVH)* [6]. We refer readers to [44] for a detailed description of deformation gradient computation.

After canonicalization, given the predicted appearance-conditioned radiance c and density σ (Sec. 3.4) for every point on each ray, the representation of the avatar is optimized by using the differentiable volumetric rendering. For each camera ray $r(t) = o + td$ with camera center o and viewing direction d , the color of the corresponding pixel can be computed by accumulating the predicting radiance and densities of the sample points along the ray:

$$\mathcal{C}(\mathbf{r}) = \int_{t_n}^{t_f} \sigma(\mathbf{r}(t)) \cdot \mathbf{c}(\mathbf{r}(t)) \cdot \mathcal{T}(t) dt \quad (2)$$

where t_n and t_f are the near and far bounds correspondingly, and $\mathcal{T}(t_n) = \exp(-\int_{t_n}^{t_f} \sigma(t) dt)$ is the accumulated transmittance along the ray from t_n to t_f , $\sigma(t)$ is the density and $c(t)$ it the radiance at position x_t .

3.3 Learned Appearance Embedding

In this section, PAV introduces a learnable appearance embedding that is essential and effective for enabling discriminative multi-appearance rendering using single unified network. To achieve this goal (See Fig. 2), a set of N learnable latent features $\mathbf{Z}=\{Z_i\}_{i=1}^N, Z_i \in \mathbb{R}^{512 \times 512 \times 16}$ are defined in UV space corresponding to N appearances of the same subject in the video collection. We impose *high-dimensional* learnable latent features to condition implicit representations for both density σ and radiance c values in the canonical space. To this end, these latent neural features are attached to the vertices of 3DMM mesh, which enables PAV to memorize local variations. For each point x sampled on a ray, we find the nearest triangle $\mathcal{T}=\{v_i\}_{i=1}^3, v_i \in \mathbb{R}^3$ (sec. 3.2) and aggregate the features $z(v_i)$ of vertices v_i from the vertices of the nearest triangle to get the appearance embedding Z^x . Then we directly use the appearance embedding to condition for appearance-dependent neural radiance field estimation (Sec. 3.4). During training, these features are randomly initialized and optimized jointly with the objectives of the entire network.

3.4 Appearance Conditioned NeRF

To model the appearance-conditioned dynamic deformable head NeRF, PAV incorporates learned appearance-embedding (Sec. 3.3) into neural radiance field in canonical space (See Fig. 2). For a point x sampled on a ray, after canonicalization, the point x_c is encoded using multi-resolution hash encoding [24]. This feature $\phi(x_c)$ is passed to multi-layer perceptron for geometric feature estimation. Following [10] and [44], we conditioned every point features $\phi(x_c)$ with the facial expressions e of frame i . Without per-frame latent code, for shared volume density, we learn implicit representation $\mathcal{F}_{density}$ using multi-layer perceptron (MLP-D in Fig. 2):

$$\mathcal{F}_{density} : (\phi(x_c), e_i) \rightarrow \sigma \quad (3)$$

A primary challenge lies in learning unique density estimations for each appearance while separating the shared density aspects from the appearance. Hence we design PAV such that appearance-based density offset is predicted from appearance embedding Z^x sampled from the nearest triangle to the point x on the mesh M_i and facial expression e_i of frame i . For density offset, we learn implicit representation \mathcal{F}_{offset} using multi-layer perceptron (MLP-O in Fig. 2):

$$\mathcal{F}_{offset} : (Z^x, e_i) \rightarrow \Delta_\sigma \quad (4)$$

Final density value $\sigma(x) = \sigma + \Delta_\sigma$ for the point x is the summation of shared density features σ and density offset Δ_σ . For emitted radiance at point x , we learn implicit representation \mathcal{F}_{color} using multi-layer perceptron (MLP-C in Fig. 2):

$$\mathcal{F}_{color} : (Z^x, d_i, \sigma(x)) \rightarrow c \quad (5)$$

where d_i is the encoded view direction of training frame i . Then, final color for the pixel is computed by using differentiable *volume rendering* (Eq. 2)

3.5 Optimization

Our model is trained on a collection of monocular RGB videos with photometric and depth losses. For photometric loss \mathcal{L}_{color} , we employ L_2 distance between the renderings C and ground-truth images I , $\mathcal{L}_{color} = \sum_i \sum_r \|C_i(r) - I_i(r)\|$, where r denotes the camera ray of each pixel and i denotes frame index. For depth loss \mathcal{L}_{depth} , we rasterize the depth of tracking mesh M_i and apply L_1 distance between this map D and the ray termination H of the volumetric rendering only for face region as FLAME model is restricted to model details like hair: $\mathcal{L}_{depth} = \sum_i \sum_r \mathcal{M}_i(r) \|H_i(r) - D_i(r)\|$, where \mathcal{M} denotes face binary mask. The total loss is defined as:

$$\mathcal{L} = \lambda_{color} \mathcal{L}_{color} + \lambda_{depth} \mathcal{L}_{depth} \quad (6)$$

where we empirically set $\lambda_{color} = 50$ and $\lambda_{depth} = 10$.

4 Experiments

In this section, we evaluate the performance of PAV using two different datasets. We perform an ablation study to validate our design choices and a comparison against the state-of-the-art monocular head avatar method in various configurations.

4.1 Dataset

Our method takes a video collection of a subject containing different appearances, poses, and expressions. Since there is no existing dataset to evaluate

learning 3D head avatar from video collections of the same subjects with various appearances, poses, and expressions, we compiled a new Video Collection (*Vid-Col*) dataset from talking face videos of *Roger Federer*, *King Charles* and *Tom Hanks* across years. This dataset consists of in-the-wild videos recorded in uncontrolled environments. In VidCol dataset, we curated videos from 5 different appearances for each subject. We randomly sampled 500 frames at 512^2 resolution from each video (or appearance) and we used a total of 2500 frames to train and test a single unified network per subject. We will share this dataset with annotations of semantic masks, tracked 3DMM models, and camera calibrations for further research in this direction.

For our experiments, we evaluate all characters and train a separate PAV model for each subject. We develop two different protocols for evaluation:

- a) *Unified Network for Novel Pose/Exp. Synthesis* This protocol evaluates the performance in terms of novel appearance, head pose, and facial expression synthesis when training a single unified network from all appearances. For training, we select 400 frames for each appearance of a subject talking in front of a camera with various head poses. It mimics a training video commonly used for monocular neural head avatar methods (Sec. 2). For testing, we use the remaining 100 frames for each appearance with unseen poses and expressions.
- b) *Separated Networks for Novel Pose/Exp. Synthesis* This protocol evaluates the novel appearance, head pose, and facial expression generalization ability of the methods when training a separate network for each appearance. We select 400 training images for each appearance of a subject talking of a camera, same as protocol-a. For testing, we choose the remaining 100 unseen pose/expression combinations with ground truth, same as protocol-a, and another 300 unseen pose/expression sets without ground truth image data.

4.2 Metrics

We report several metrics to evaluate the quality of synthesized images as well as the ablation study of our method. For synthesized images with provided ground-truth, we use Learned Perceptual Patch Similarity (LPIPS) [41], Deep Image Structure and Texture Similarity (DISTTS) [8], Structural Similarity Index (SSIM) [36], Peak Signal-to-Noise Ratio (PSNR) and Fréchet Inception Distance (FID) [15]. When there is an absence of ground truth, we adopt FID metric for quantitative comparison.

4.3 Implementation Details

We optimize Eq. 6 using Adam optimizer [17] with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. We set the learning rate to 2.5×10^{-3} for $F_{density}$, F_{color} and Z , and to 2.5×10^{-4} for F_{offset} . We sample 1024 points along each ray for rendering. Multi-resolution hash encoding used in the method is set to 16 *number of levels*, 18 *hash map size*, and 2048 *finest resolution*. The optimization takes 15K iterations

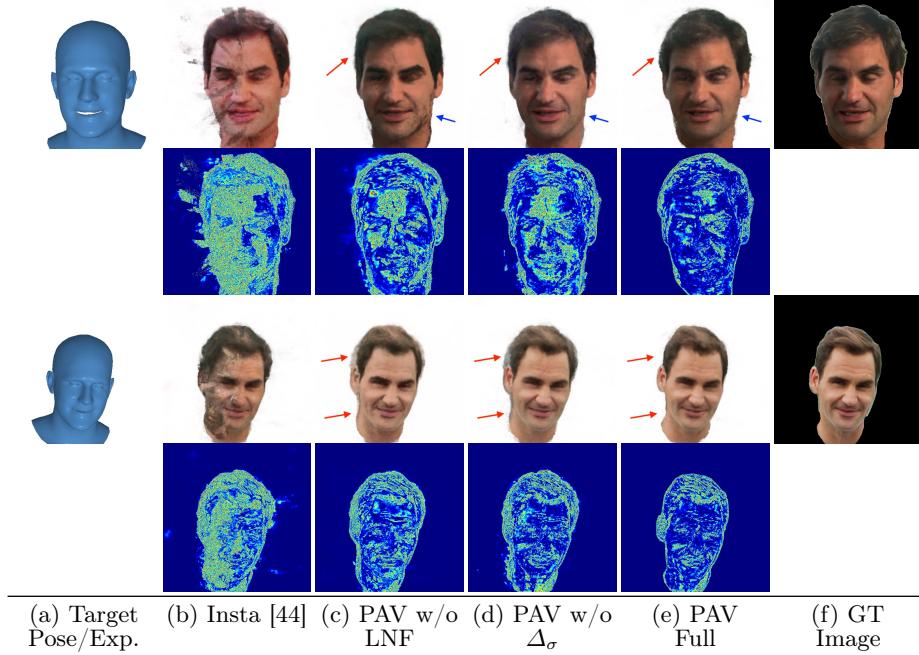


Fig. 3: Qualitative Ablation Study on VidCol Dataset. This figure shows synthesized images and pixel error maps against ground-truth. Please note that brighter pixel colors denote higher pixel errors. Without appearance embedding with latent neural features and appearance-conditioned density offset (a), the model fails to learn appearance-based distinct details. The latent neural features (LNF) (c) resolve the issue of high-fidelity rendering of texture (blue arrow). The appearance-conditioned density offset field (d) allows the model to learn more accurate geometry and texture that corresponds to the face, hair, and neck regions (red arrow)

to converge when training network with all appearances of single subjects. More details on the network are provided in the supplementary document.

4.4 Ablation Study

We evaluate different design choices of PAV in Tab. 1 and Fig. 3. In this study, we use protocol-a of the VidCol dataset for all experiments. We first report the results of the final model which includes *latent neural features (LNF)* (Sec. 3.3) and density offset Δ_σ (Sec. 3.4). The full model achieves PSNR score of 25.15 and SSIM score of 0.91 for image synthesis. If we remove the density offset from our model while employing appearance-conditioned radiance estimation, PSNR and SSIM scores for image synthesis decrease to 24.14 and 0.90, respectively. This indicates that without appearance-conditioned density offset Δ_σ , density estimation MLP ($\mathcal{F}_{density}$ in Sec. 3.4) can not learn to synthesize appearance-specific geometry and texture details, thus rendering fails there as well. Even

	PSNR \uparrow	SSIM \uparrow	DISTS \downarrow	LPIPS \downarrow	FID \downarrow
Insta [44]	21.03	0.849	0.185	0.169	104.28
PAV without LNF	24.19	0.900	0.145	0.119	54.39
PAV without Δ_σ	24.14	0.906	0.134	0.108	47.71
PAV (Full Model)	25.15	0.910	0.133	0.107	46.67

Table 1: Ablation Study: We evaluate the impact of different components of the proposed method. See Fig. 3 for qualitative comparison.

though the model w/o Δ_σ synthesise images that look on par with ground truth in terms of texture, it generates severe artifacts which can be seen in pixel error maps. An example of this behavior can also be seen in Fig. 3 (w/o Δ_σ). Rendering failures on face, head, and neck regions are indicated by **red arrow** and error map.

Next, we evaluate the impact of *latent neural features* (LNF) as an *appearance embedding* on PAV. In this experiment, to condition the canonical MLPs in PAV, inspired by "Nerf In-the-wild" [22, 37], we adopt Generative Latent Optimization (GLO) for appearance embedding. Specifically, this approach defines a single real-valued appearance embedding vector l for each appearance in the dataset. To this end, we employ a pre-trained encoder from in-the-wild face detector RetinaFace [7] for GLO. During training, if we replace $\mathbf{Z}=\{Z_i\}_{i=1}^N, Z_i \in \mathbb{R}^{512 \times 512 \times 16}$ with $\mathbf{L}=\{l_i\}_{i=1}^N, l_i \in \mathbb{R}^{512 \times 1}$ for N appearances of each subject, and pass it to appearance conditioned NeRF (See Sec. 3.4), PSNR and SSIM scores for image synthesis decrease to 24.19 and 0.90, respectively. This is also evident from the qualitative results shown in Fig. 3 (w/o LNF), indicating that latent neural features help in improved image synthesis quality with their local features attached to 3D geometry. Rendering failures on head, neck, and skin texture are shown by **red arrow**, **blue arrow** and error maps in Fig. 3. Finally, as discussed in Sec. 3, we keep the network design as we explain after the ablation study.

4.5 Comparison

Since our method is the first work that learns dynamic deformable head NeRF from multiple appearances of the same subjects using a unified network, we ourselves design some baseline comparisons in this section. First, we compare our method against Insta [44] using protocol-a (Sec. 4.1) of the VidCol dataset. Both methods are trained using the frames sampled from all appearances of the same subject. For testing, both methods synthesize images for target head pose and expression. The results are summarized in Tab. 2 and Fig. 4. For protocol-a, our method clearly outperforms the baseline for final image synthesis results across all metrics. It should be noted that PAV framework jointly learns all different appearances via one-time training using a unified network and it is able to preserve appearance-specific details for novel poses and expressions. On the other hand, as shown in Fig. 4, Insta can not disentangle appearance from the

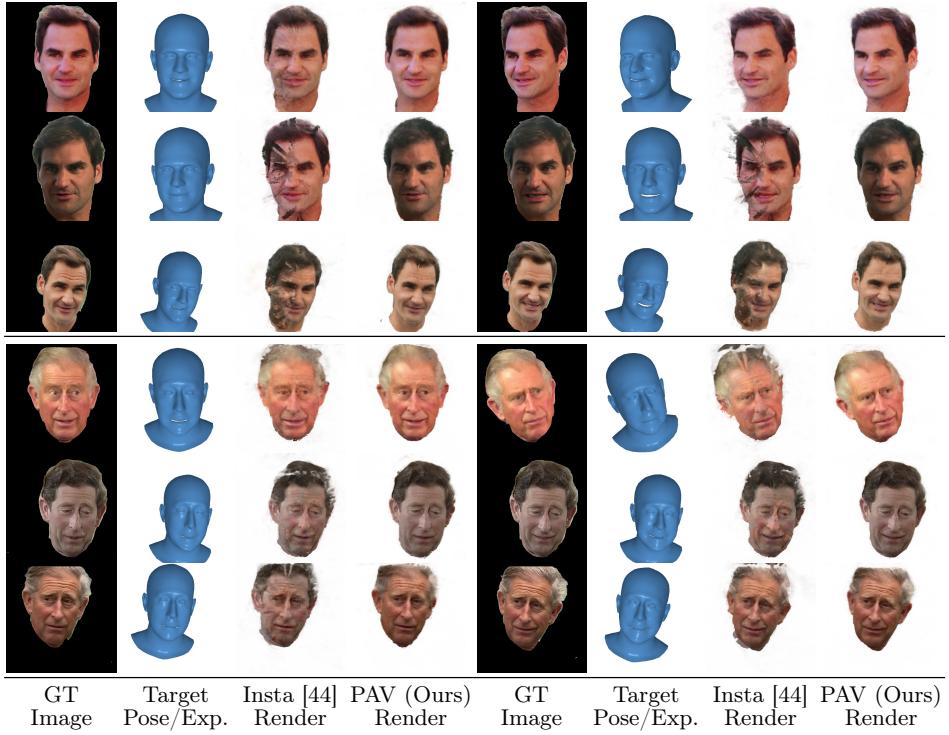


Fig. 4: Qualitative comparison of our method (PAV) against baseline method (Insta [44]). Columns {2,6} show target head poses and facial expressions for synthesized images at columns {3,4,7,8}.

neural radiance field, which results in a mixture of appearances in synthesized images and inaccurate geometry predictions. This is the limitation of Insta we are addressing in the paper.

Second, we train separate Insta [44] models for each appearance of the same subject using protocol-b while training a single unified model for our method using the frames sampled from all appearances of the same subject. For testing, we use the same test data from protocol-b for both PAV and Insta [44] and synthesize images for novel poses and expressions. Note that, we added more novel expressions and poses in protocol-b (Sec. 4.1) to test the generalization of our method under a unified training schema, which results in increasing FID score for both methods in Tab. 3. Even though the results show that artifacts (in Fig. 4) of the Insta renders are decreased in Fig. 5, our method demonstrates better-rendering results (Fig. 5) and lower FID score (Tab. 3). This indicates that PAV’s single unified model helps in learning generic neural radiance field representation across all appearances of the same subject while preserving appearance-based details. Please note that we are demonstrating Insta’s performance under a challenging pose/expression in Fig. 5 which can also be seen in

Method	Video Collection 1					Video Collection 2					
	PSNR ↑	SSIM ↑	DISTS ↓	LPIPS ↓	FID ↓		PSNR ↑	SSIM ↑	DISTS ↓	LPIPS ↓	FID ↓
Protocol (a): Novel Head Pose and Expression Synthesis											
Insta [44]	21.03	0.84	0.185	0.169	104.28		23.67	0.86	0.187	0.163	98.87
PAV (Ours)	25.15	0.91	0.133	0.107	46.67		27.09	0.89	0.144	0.120	53.91

Table 2: Comparison of PAV against Insta [44]. See Fig. 4 for qualitative comparison. Video Collection 1 and 2 refer to *Federer* and *King Charles* splits of *VidCol* dataset, respectively.

the original paper results under extreme poses/expressions. In Fig. 5, the reference image is for photoconsistency check only since there is no ground truth image for that pose/expression. Finally, we provide additional qualitative results of avatars animated in the supplementary video. We also compared PAV against Insta for model efficiency using protocol-b dataset and recorded values in Tab. 3. Although we train our model with multiple appearances jointly, PAV is more efficient than Insta in terms of memory footprint and training time

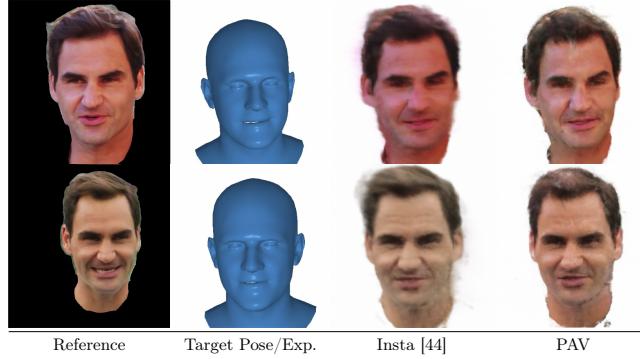


Fig. 5: Qualitative comparison of PAV against single-appearance trained Insta [44]. Please note that the reference image is for photoconsistency check only since there is no ground truth image for that pose/expression

5 Discussion

We presented PAV which is a novel method for learning personalized head avatars from monocular and in-the-wild video collections of a person. We demonstrated that PAV can model a 3D head avatar from an unstructured video collection of a subject in a single unified network while disentangling face appearances from the radiance field. PAV can be trained simultaneously for multiple appearances of the same subject, allowing novel pose and expression generalization and leading to a superior rendering quality of our method compared to the baseline.

	Train Time Minutes	Number of Param. Million	Model Size (MB)	FID ↓
Multiple Insta [44]	135	79.7	1932	109.22
PAV (Ours)	45	26.8	650	91.01

Table 3: Efficiency and performance comparison between PAV and Insta [44] trained for each appearance (Single Appearance Training) in dataset collection 1. FID score is computed from synthesized images under novel head poses and expressions where there are no existing ground-truth images.

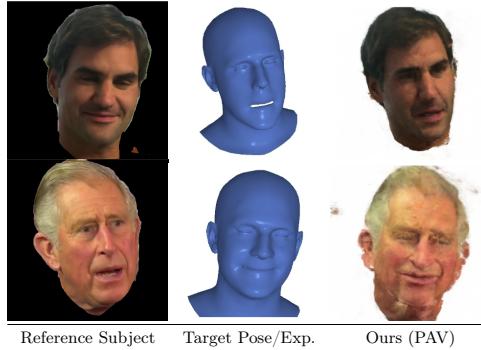


Fig. 6: Multi-identity and multi-appearance extension of PAV. The learned appearance (middle) rendered in a new viewpoint (left) does not match the visual quality of input images (right). Please note that the reference image is for photoconsistency check only since there is no ground truth image for that pose/expression.

Limitation for Multi-Identity. We extended our method for multiple identities and multiple appearances for each identity. To this end, we train a single unified PAV model using frames sampled from all appearances of *Federer* and *King* subjects from *VidCol* dataset with shared canonical space. We evaluate novel head pose and facial expression rendering of this model and Fig. 6 illustrates the results. Even though this model produces reasonable results, synthesized images demonstrate a lack of subject-specific shape and appearance details with rendering artifacts. In the paper, we focus on multiple appearances of a single subject. Scaling up this model to a wide range of subjects and exploring generalized dynamic deformable NeRF would be interesting future work.

Ethical Considerations. Although the rapid progress on digital human avatars facilitates numerous downstream applications, it also raises concerns about ethical misuse. On the synthesis side, it would be ideal to actively use watermarking technologies and avoid driving avatars with different identities. In this work, we aim to faithfully reproduce images of a person with the capability of rendering unseen poses and expressions and switching appearance within their own set of appearances. The work does not intend to create animations that didn't happen.

References

1. Athar, S., Xu, Z., Sunkavalli, K., Shechtman, E., Shu, Z.: Rignerf: Fully controllable neural 3d portraits. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 20364–20373 (2022)
2. Averbuch-Elor, H., Cohen-Or, D., Kopf, J., Cohen, M.F.: Bringing portraits to life. ACM transactions on graphics (TOG) **36**(6), 1–13 (2017)
3. Bai, Z., Tan, F., Huang, Z., Sarkar, K., Tang, D., Qiu, D., Meka, A., Du, R., Dou, M., Orts-Escalano, S., et al.: Learning personalized high quality volumetric head avatars from monocular rgb videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16890–16900 (2023)
4. Bühler, M.C., Meka, A., Li, G., Beeler, T., Hilliges, O.: Varitex: Variational neural face textures. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13890–13899 (2021)
5. Chen, C., O'Toole, M., Bharaj, G., Garrido, P.: Implicit neural head synthesis via controllable local deformation fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 416–426 (2023)
6. Clark, J.H.: Hierarchical geometric models for visible surface algorithms. Communications of the ACM **19**(10), 547–554 (1976)
7. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641 (2019)
8. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image Quality Assessment: Unifying Structure and Texture Similarity. TPAMI (2022)
9. Egger, B., Smith, W.A., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., et al.: 3d morphable face models—past, present, and future. ACM Transactions on Graphics (ToG) **39**(5), 1–38 (2020)
10. Gafni, G., Thies, J., Zollhöfer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8649–8658 (June 2021)
11. Gao, X., Zhong, C., Xiang, J., Hong, Y., Guo, Y., Zhang, J.: Reconstructing personalized semantic facial nerf models from monocular video. ACM Transactions on Graphics (TOG) **41**(6), 1–12 (2022)
12. Garrido, P., Valgaerts, L., Rehmsen, O., Thormahlen, T., Perez, P., Theobalt, C.: Automatic face reenactment. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4217–4224 (2014)
13. Garrido, P., Valgaerts, L., Sarmadi, H., Steiner, I., Varanasi, K., Perez, P., Theobalt, C.: Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In: Computer graphics forum. vol. 34, pp. 193–204. Wiley Online Library (2015)
14. Grassal, P.W., Prinzler, M., Leistner, T., Rother, C., Nießner, M., Thies, J.: Neural head avatars from monocular rgb videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18653–18664 (2022)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
16. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. ACM transactions on graphics (TOG) **37**(4), 1–14 (2018)

17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Kuang, Z., Olszewski, K., Chai, M., Huang, Z., Achlioptas, P., Tulyakov, S.: Neroic: Neural rendering of objects from online image collections. ACM Transactions on Graphics (TOG) **41**(4), 1–12 (2022)
19. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6), 194:1–194:17 (2017), <https://doi.org/10.1145/3130800.3130813>
20. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control. ACM transactions on graphics (TOG) **40**(6), 1–16 (2021)
21. Mallya, A., Wang, T.C., Liu, M.Y.: Implicit warping for animation with image sets. Advances in Neural Information Processing Systems **35**, 22438–22450 (2022)
22. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7210–7219 (2021)
23. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
24. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
25. Nirkin, Y., Keller, Y., Hassner, T.: Fsgan: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7184–7193 (2019)
26. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. arXiv preprint arXiv:2106.13228 (2021)
27. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14314–14323 (2021)
28. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
29. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in neural information processing systems **32** (2019)
30. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-nerf: Scalable large scene neural view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8248–8258 (2022)
31. Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., Martin-Brualla, R., Simon, T., Saragih, J., Nießner, M., et al.: State of the art on neural rendering. In: Computer Graphics Forum. vol. 39, pp. 701–727. Wiley Online Library (2020)
32. Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Yifan, W., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., et al.: Advances in neural rendering. In: Computer Graphics Forum. vol. 41, pp. 703–735. Wiley Online Library (2022)

33. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)* **38**(4), 1–12 (2019)
34. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2387–2395 (2016)
35. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. *TIP* (2004)
37. Weng, C.Y., Srinivasan, P.P., Curless, B., Kemelmacher-Shlizerman, I.: Personerf: Personalized reconstruction from photo collections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 524–533 (2023)
38. Yang, B., Bao, C., Zeng, J., Bao, H., Zhang, Y., Cui, Z., Zhang, G.: Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In: European Conference on Computer Vision. pp. 597–614. Springer (2022)
39. Zakharov, E., Ivakhnenko, A., Shysheya, A., Lempitsky, V.: Fast bi-layer neural synthesis of one-shot realistic head avatars. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. pp. 524–540. Springer (2020)
40. Zhang, B., Qi, C., Zhang, P., Zhang, B., Wu, H., Chen, D., Chen, Q., Wang, Y., Wen, F.: Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22096–22105 (June 2023)
41. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
42. Zheng, Y., Abrevaya, V.F., Bühler, M.C., Chen, X., Black, M.J., Hilliges, O.: Im avatar: Implicit morphable head avatars from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13545–13555 (2022)
43. Zielonka, W., Bolkart, T., Thies, J.: Towards metrical reconstruction of human faces. In: European Conference on Computer Vision. pp. 250–269. Springer (2022)
44. Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4574–4584 (2023)