Pixel-aligned Volumetric Avatars

Amit Raj[1] Michael Zollhöfer [2] Tomas Simon [2] Jason Saragih [2]
Shunsuke Saito [2] James Hays [1] Stephen Lombardi [2]

[1] Georgia Institute of Technology [2] Facebook Reality Labs Research

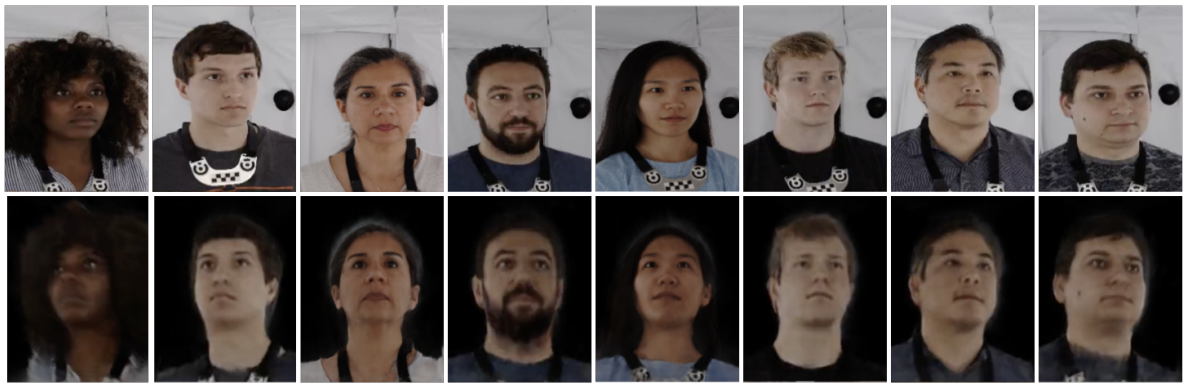

Figure 1: We present a novel approach for the prediction of volumetric avatars of human heads from a small number of example views. Our model enables view synthesis for unseen identities and is able to generate faithful facial expressions.

Abstract

Acquisition and rendering of photo-realistic human heads is a highly challenging research problem of particular importance for virtual telepresence. Currently, the highest quality is achieved by volumetric *approaches trained in a person-specific manner on multi-view data. These models better represent fine structure, such as hair, compared to simpler mesh-based models. Volumetric models typically employ a global code to represent facial expressions, such that they can be driven by a small set of animation parameters. While such architectures achieve impressive rendering quality, they can not easily be extended to the multi-identity setting. In this paper, we devise a novel approach for predicting volumetric avatars of the human head given just a small number of inputs. We enable generalization across identities by a novel parameterization that combines neural radiance fields with local, pixel-aligned features extracted directly from the inputs, thus side-stepping the need for very deep or complex networks. Our approach is trained in an end-to-end manner solely based on a photometric re-rendering loss without requiring explicit 3D supervision. We demonstrate that our approach outperforms the existing state of the art in terms of quality and is able to generate faithful facial expressions in a multi-identity setting.*

1. Introduction

The acquisition and rendering of photo-realistic human heads is a highly challenging research problem with high significance for virtual telepresence applications. Human heads are challenging to model and render due to their complex geometry and appearance properties: sub-surface scattering of skin, fine-scale surface detail, thin-structured hair, and the human eyes as well as the teeth are both specular and translucent. Most existing approaches require complex and expensive multi-view capture rigs (with up to hundreds of cameras) to reconstruct even a person-specific model of a human head.

Currently, the highest-quality approaches are those that employ *volumetric* models rather than a textured mesh, since they can better learn to represent fine structures on the face like hair, which is critical to achieving a photo-realistic appearance. These volumetric models [13] typically employ a global code to represent facial expressions or only work for static scenes [16, 9]. While such architectures achieve impressive rendering quality, they can not easily be adapted to a multi-identity setting. A global code, as is used to control expression, is not sufficient for modeling identity variation across subjects. There has been significant progress of late in using implicit models to represent scenes and objects. These models have the advantage

that the scene is represented as a parametric function in a continuous space, which allows for fine-grained inference of geometry and texture [22]. But these approaches can not model view-dependent effects and it is challenging to represent for example hair with a textured surface. The approach of Sitzmann et al. [27] can generalize across objects, but only at low resolutions and can only handle purely Lambertian surfaces, which is not sufficient for human heads. Despite the recent success and advantages of such scene representation approaches, there are several limitations. In particular, most of the above methods train a network to model only a single scene or object. Methods which can generate multiple objects are typically limited in terms of quality and resolution of the predicted texture and geometry.

We present pixel-aligned volumetric avatars (PVA), a novel framework for the estimation of a volumetric 3D avatar from only a few input images of a human head. Our approach is able to generalize to unseen identities at test time. Methods such as Scene Representation Networks (SRNs) [25], which generate a set of weights from a global image encoding (i.e., a single latent code vector per image), have difficulty generalizing to local changes (e.g., facial expressions) and fail to recover high-frequency details even when these are visible in the input images. This is because the global latent code summarizes information in the image and must discard some information to generate a compact encoding of the data. To improve generalization across identities, we instead parameterize the volumetric model via local, pixel-aligned features extracted from the input images.

We show that our model can synthesize novel views for unseen identities and expressions while preserving high frequency details in the rendered avatar. To summarize, our contribution are:

- We introduce a novel pixel-aligned radiance field that predicts implicit shape and appearance from a sparse set of posed images.
- Our model generalizes to unseen identities and expressions at test time.
- We demonstrate state of the art performance on novel view synthesis compared to recent approaches.

2. Related Work

Generating avatars from images has a long history in computer vision and graphics. Traditional methods employ mesh-based representations and physics-inspired models of how faces deform and interact with light, while more recent approaches employ deep learning to overcome some of the limitations of classical techniques. We discuss several classes of methods below and compare them to ours.

Mesh-based Approaches Active Appearance Models (AAMs) was among the first face models capable of modeling facial expressions, although it was originally used as a statistical joint shape and appearance model for human faces [4], and later extended to 3D faces [2]. Deep Appearance Models [12, 17] create a 3D morphable model using deep networks to create an extremely high-quality and driveable face model. However mesh-based methods struggle with rendering thin structures like hair, which are critical for realistic human face rendering. Mesh-based methods have been extended in a number of ways to improve quality and expressiveness, though they typically share similar disadvantages. Notably, mesh-based models require a fixed topology, which poses problems for modeling hair, which can vary dramatically from one person to another. Furthermore, mesh-based methods have hard triangle boundaries which can look unpleasant for soft features. Finally, optimizing meshes to match the appearance of arbitrary shapes is still a difficult problem. Efforts in differentiable rasterizers [3, 11, 8, 6] have shown impressive results in generating meshes from single and multi-view images without 3D supervision, but the generated meshes usually have restrictions in terms of topology and fail to capture high frequency details. Furthermore, they are limited in terms of the textures that can be represented. In contrast, our method is able to capture arbitrary topology (as seen in expressions and hairstyles) and captures high frequency texture details better, since it is able to use pixel-level information more efficiently.

Image-based Methods Recently, there has been a great deal of progress in high-quality controllable face synthesis [7, 30, 1]. However, these image-based methods work with mostly frontal faces and have difficulty explicitly controlling the viewpoint and expression of the synthesized images. Without giving the network a notion of 3D space, it is difficult for the methods to generalize without many training images. StyleRig [28] enables parameteric control of StyleGAN generated imagery. However, the results are not multi-view consistent and the approach does not work on real images.

Voxel-based Methods Methods such as [33, 19, 26] learn an intermediate 3D voxel grid of features and a 3D-2D projection operation to synthesize images. Transformable bottleneck networks [21] present a method that learns a bottleneck of 3D features that can be manipulated directly to enable a variety of applications. However, the primary problem with these voxel-based approaches are their inability to scale to higher resolution due to memory restrictions. We eschew the problem of capacity by learning a multi-layer perceptron (MLP) that directly translates 3D locations and pixel-aligned features to color and occupancy.

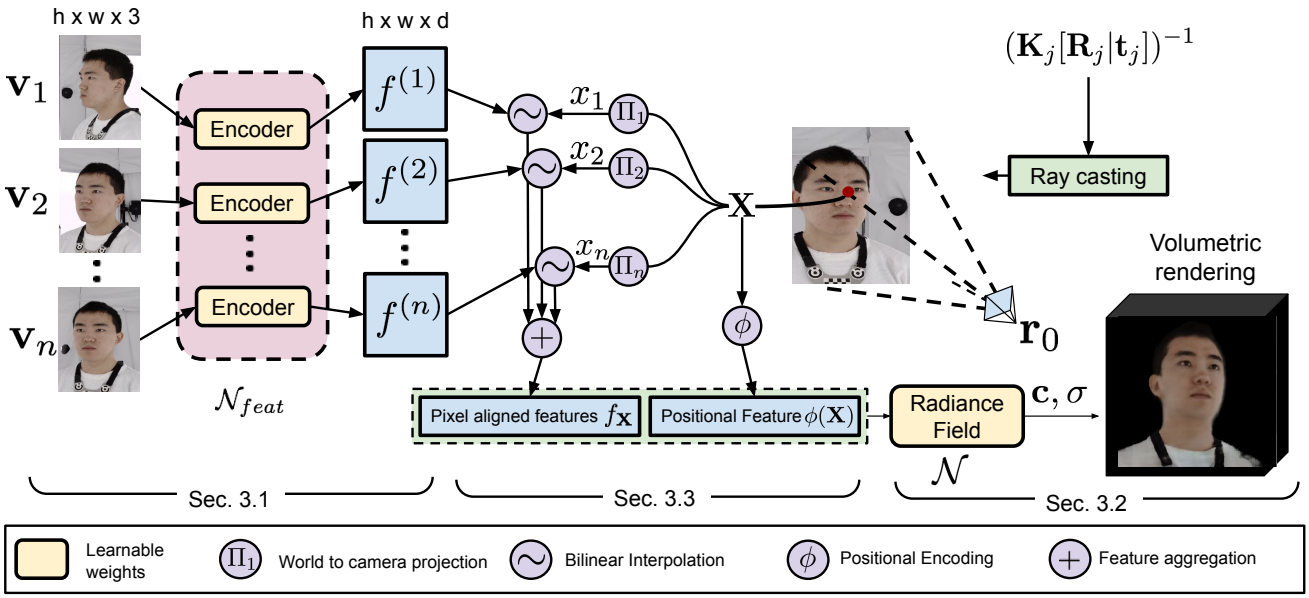

Figure 2: Overview of the proposed approach. Given a target viewpoint and a set of conditioning images, our novel approach employs local, pixel-aligned features that are extracted from the inputs to condition a multi-identity neural radiance field. Volume rendering is employed to generate an image of the subject from the target viewpoint.

Implicit Methods Works such as [31] use explicit 3D information during training. PIFu/PIFuHD [22, 23] models human bodies with an implicit function evaluated at the depth of a point. It is capable of rendering human bodies with high quality. A key insight is to use pixel-aligned features to retain high-frequency detail. We leverage this insight, but our method does not require 3D supervision. Scene Representation Networks (SRNs) [27] model scenes with a learned SDF and do not require 3D supervision. We do not assume a well-defined surface through an SDF but rather a semi-transparent representation that can better model hair and thin structures. The authors of [18, 5] learn an implicit representation of geometry from natural images in an unsupervised manner to allow novel view synthesis. These methods are limited in the degree of multi-view consistency that can be achieved. TextureFields [20] learn to transfer textures from an exemplar image to a source mesh to allow novel view synthesis. We eschew the need for a mesh at inference time by learning an implicit representation of geometry.

Neural Rendering Many neural rendering models have been proposed recently that better represent thin structures, like hair and clothes. Neural volumes [13] and NeRF [16] are two recently introduced methods that model objects with a semitransparent volume and have shown the ability to model thin structures well. Neural Volumes can also model dynamic scenes. NeRF-W [14] extends the work of [16] to a conditional setting to models scenes under different lighting with same underlying geometry. However, these methods fail to generalize to novel identities. Inspired by insights from NeRF and PIFu, we demonstrate a framework that handles multiple identities by relying on pixel-aligned features. GRAF [24] learns a conditional radiance field in an unsupervised manner by disentangling a global shape and appearance code which limits its ability to model local shape and texture deformations. Other works focus on speeding up NeRF using a sparse Octree structure [10]. We refer the readers to the recent STAR of Tewari et al. [29] for an in-depth treatment of recent neural rendering methods.

3. Approach

We present a Pixel-aligned Volumetric avatars(PVA). An implicit model of faces that is learned from a multi-view image collection, see Fig. 2. Our model can generate novel views of unseen identities from one or more example images. The framework consists of two main components. The first is a shallow convolutional encoder-decoder ($\mathcal{N}_{feat}$) network that takes as input one or more images ($\mathbf{v}_i$) of a person from a known viewpoint $\{\mathbf{K}_i, [\mathbf{R}|\mathbf{t}]_i\}$ and produces pixel-aligned feature maps $f^{(i)}$. The second component is a radiance field network ($\mathcal{N}$) that converts 3D location and pixel-aligned features to color and opacity. To render the radiance field, we march along the camera ray of each pixel in the target view j, defined by $\{\mathbf{K}_j, [\mathbf{R}|\mathbf{t}]_j\}$, accumulating the color and occupancy produced by $\mathcal{N}$ at each

point. We train our approach based on a multi-identity training corpus using gradient descent. To this end, we minimize the L_2 loss between predicted images and the corresponding ground truth.

3.1. Pixel-aligned Radiance Fields

We employ a pixel-aligned scene representation modeled as a neural network. Concretely, for a conditioning view $\mathbf{v}_i \in \mathbb{R}^{h\times w\times 3}$ we define functions

$$f^{(i)} = \mathcal{N}_{\text{feat}}(\mathbf{v}_i) \tag{1}$$

$$\{\mathbf{c}, \sigma\} = \mathcal{N}(\phi(\mathbf{X}), f_{\mathbf{X}}) \tag{2}$$

where $\phi(\mathbf{X}) : \mathbb{R}^3 \rightarrow \mathbb{R}^{6\times l}$ is the positional encoding of $\mathbf{X} \in \mathbb{R}^3$ as in [16] with $2 \times l$ different basis functions, $f^{(i)} \in \mathbb{R}^{h\times w\times d}$ is the feature map of $\mathbf{v}_i$, d the number of feature channels, h and w are image height and width, and $f_{\mathbf{X}} \in \mathbb{R}^{d'}$ is the aggregated image feature associated with the point $\mathbf{X}$ as explained in the next section. For each feature map $f^{(i)}$, we obtain $f^{(i)}_{\mathbf{X}} \in \mathbb{R}^d$ by projecting 3D point $\mathbf{X}$ along the ray using camera intrinsic and extrinsic parameters $\mathbf{K}, \mathbf{R}, \mathbf{t}$ of that particular viewpoint,

$$x_i = \Pi(\mathbf{X}; \mathbf{K}_i\,[\mathbf{R}|\mathbf{t}]_i), \tag{3}$$

$$f^{(i)}_{\mathbf{X}} = \mathcal{F}(f^{(i)}; x_i) \tag{4}$$

where Π is a perspective projection function to camera pixel coordinates, and $\mathcal{F}(f, x)$ is the bilinear interpolation of f at pixel location x.

3.2. Volume Rendering

For each given training image $\mathbf{v}_j$ with camera intrinsics $\mathbf{K}_j$ and rotation and translation $\mathbf{R}_j, \mathbf{t}_j$, the predicted color of a pixel $p \in \mathbb{R}^2$ for a given viewpoint in the focal plane of the camera and center $\mathbf{r}_0 \in \mathbb{R}^3$ is obtained by marching rays into the scene using the camera-to-world projection matrix, $\mathbf{P}^{-1} = [\mathbf{R}_i|\mathbf{t}_i]^{-1}\mathbf{K}_i^{-1}$ with the direction of the rays given by,

$$\mathbf{d} = \frac{\mathbf{P}^{-1}p - \mathbf{r}_0}{\|\mathbf{P}^{-1}p - \mathbf{r}_0\|}. \tag{5}$$

Note that in order to help the network focus its capacity on modeling the content of the scene, all camera extrinsics are relative to the computed head pose, which is found via traditional head registration.

We then accumulate the radiance and opacity along the ray $\mathbf{r}(t) = \mathbf{r}_0 + t\mathbf{d}$ for $t \in [t_{\text{near}}, t_{\text{far}}]$ as defined in NeRF [15] as follows:

$$\mathbf{I}_{rgb}(p) = \int_{t_{\text{near}}}^{t_{\text{far}}} \mathbf{T}(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt \tag{6}$$

where,

$$\mathbf{T}(t) = \exp\left(-\int_{t_{\text{near}}}^{t} \sigma(\mathbf{r}(s))ds\right) \tag{7}$$

In practice we uniformly sample a set of n_s points $t \sim [t_{near}, t_{far}]$. We set $\mathbf{X} = \mathbf{r}(t)$ and use the quadrature rule to approximate the integral. We also define $\mathbf{I}_\alpha(p)$ as,

$$\mathbf{I}_\alpha(p) = \sum_{i=1}^{n_s} \alpha_i \prod_{j=1}^{i}(1 - \alpha_j) \tag{8}$$

where $\alpha_i = 1 - \exp(-\delta_i\sigma_i)$ with δ_i being the distance between the $i+1$-th and i-th sample point along the ray.

3.3. Multi-view Feature Aggregation

A critical component of our method is how to fuse pixel-aligned features $f^{(i)}_{\mathbf{X}}$ from multiple images to help the network best use this information.

3.3.1 Fixed number of conditioning views

In a multi-view setting with known camera viewpoints and a fixed number of conditioning views we can aggregate the features by simple concatenation [13]. Concretely, for n conditioning images $\{\mathbf{v}_i\}_{i=1}^n$ with corresponding rotation and translation matrices given by $\{\mathbf{R}_i\}_{i=1}^n$ and $\{\mathbf{t}_i\}_{i=1}^n$. We obtain n features $\{f^{(i)}_{\mathbf{X}}\}_{i=1}^n$ for each point $\mathbf{X}$ as in Eq. 3 and generate the final feature as follows,

$$f_{\mathbf{X}} = [f^{(1)}_{\mathbf{X}} \bigoplus f^{(2)}_{\mathbf{X}} \ldots \bigoplus f^{(n)}_{\mathbf{X}}]$$

where $\bigoplus$ represents concatenation along the depth dimension. This preserves feature information from all the viewpoints, leaving the MLP to figure out how to best combine and employ the conditioning information.

3.3.2 Variable number of conditioning views

The more interesting use case is to make the model agnostic to viewpoint and number of conditioning views. Simple concatenation as above is insufficient in this case, since we do not know the number of conditioning views a priori, leading to different feature dimensions during inference time. To summarize features for a multi-view setting we need a permutation invariant function $\mathcal{G} : \mathcal{R}^{n\times d} \rightarrow \mathcal{R}^d$ such that for any permutation ψ,

$$\mathcal{G}([f^{(1)}, f^{(2)}, ..., f^{(n)}]) = \mathcal{G}([f^{\psi(1)}, f^{\psi(2)}, ..., f^{\psi(n)}]).$$

A simple permutation invariant function for feature aggregation is the mean of the sampled features (as employed in PIFu [22]). This is a reasonable aggregation procedure when we have depth information during training. However, since we have inherent depth ambiguity (since the points are projected onto the feature image before sampling) we find that this kind of aggregation produces artifacts. Fig. 9 shows an example of this behavior.

This simple mean of image features does not consider camera information, which may help the network use the conditioning information more effectively. To inject viewpoint information into the feature, we learn another network $\mathcal{N}_{cf} : \mathcal{R}^{d+7} \rightarrow R^{d'}$ that takes the feature vector and the camera information ($\mathbf{c}_i$) and produces a *camera-summarized* feature vector. These modified vectors are then averaged for all conditioning views as follows

$$f'^{(i)}_{\mathbf{X}} = \mathcal{N}_{cf}(f^{(i)}_{\mathbf{X}}, \mathbf{c}_i) \tag{9}$$

$$f_{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} f'^{(i)}_{\mathbf{X}} \tag{10}$$

The advantage of this approach is that the camera-summarized features can take likely occlusions into account before the feature average is performed. The camera information is encoded as a 4D rotation quaternion and 3D camera position.

3.4. Background Model

To avoid learning parts of the background in the scene representation, we define a background estimation network: $\mathcal{N}_{bg} : \mathcal{R}^{n_c} :\rightarrow \mathcal{R}^{h\times w\times 3}$ to learn a per-camera fixed background. Particularly, we predict the final image pixels as

$$\mathbf{I}_p = \mathbf{I}_{rgb} + (1 - \mathbf{I}_\alpha)\mathbf{I}_{bg} \tag{11}$$

with $\mathbf{I}_{bg} = \bar{\mathbf{I}}_{bg} + \mathcal{N}_{bg}(C_i)$ for camera C_i where $\bar{\mathbf{I}}_{bg}$ is an initial estimate of the background extracted using inpainting. These inpainted backgrounds are often noisy leading to 'halo' effects around the head of the person (Fig. 7). Our background estimation model learns the residual to the inpainted background. This has the advantage of not needing a high capacity network to account for the background.

3.5. Color Correction Model

The different camera sensors have a slightly different response to the same incident radiance despite the fact that they are the same camera model. If nothing is done to address this, the intensity differences end up baked into the scene representation $\mathcal{N}$, which will cause the image to unnaturally brighten or darken from certain view points. To address this, we learn a per-camera bias and gain value. This allows the system to have an 'easier' way to explain this variation in the data.

3.6. Loss Function

For ground truth target images $\mathbf{v}_j$, we train both the radiance field and feature extraction network using a simple photo-metric reconstruction loss:

$$\mathcal{L}_{\text{photo}} = \|\mathbf{I}_{p_j} - \mathbf{v}_j\|_2 \,.$$

Note, our approach is trained in an end-to-end manner solely based on this 2D re-rendering loss without requiring explicit 3D supervision.

4. Experiments

We describe the setup used to capture the training data, describe the baselines used for comparison, and perform quantitative as well as qualitative comparisons.

4.1. Training Setup

Our capture setup consists of 53 cameras positioned around the subject. For each subject, we record a set of 30 expressions with a hair-cap. And a neutral expression with no hair-cap. Each frame is fit with a 3D face model including rigid head pose which we use to center the volume between different identities and expressions. We do not use any of the mesh information during training. We train our network on 50 subjects using 40 viewpoints and test on held out viewpoints. Additionally, for the expression-based model we train our network on 25 expressions and test on the remaining expressions. During training, we divide each target image into a 16×16 grid, and randomly sample a ray from each grid location for a total of 256 rays per training image. Further, we sample $n_s = 128$ points along the ray while clamping the sample points to lie in a unit volume cube. We train our model with a batch-size of 4. Our model takes around 24 hours to converge on 4 Nvidia Tesla V100s.

4.2. Baselines

In the following, we introduce the baselines we employ for the qualitative and quantitative comparisons.

Reality Capture: Is a commercially available software based on classical structure-from-motion (SFM) and multi-view stereo (MVS), that reconstructs a 3d model from a set of captured images.

Neural Volumes: Neural volumes (NV) is a voxel-based inference method that globally encodes dynamic images of a scene and decodes a voxel grid and a warp field that represents the scene.

cNeRF: We trained a variant of NeRF with global identity conditioning (cNeRF). Particularly, we employ a VGG-network to extract a single 64D feature vector for each training identity and condition NeRF additionally on this input.

4.3. Qualitative Comparisons

We demonstrate novel view synthesis of unseen identities using our pixel aligned radiance fields, see Fig. 3. As can be seen, given only two views as input, our approach predicts volumetric avatars that can be viewed from a large number of novel viewpoints.

We also compare our method against three baselines that can handle *unseen identities* and *do not use explicit 3D supervision* for training in Fig. 4. In all baselines and for our approach, we employ only two images of the novel identities as input to compute the reconstruction. As can be seen, our approach outperforms all baselines in terms

Figure 3: We demonstrate novel view synthesis of unseen identities using pixel aligned radiance fields. All volumetric avatars were computed given only two views as input.

Figure 4: We compare our approach against three baselines: Reality Capture (a), Neural Volumes (b), Globally conditioned NeRF (c). We also show our result (d) and the ground truth identity (e). As can be seen, our approach outperforms the other methods in terms of completeness and level of reconstructed detail by a large margin.

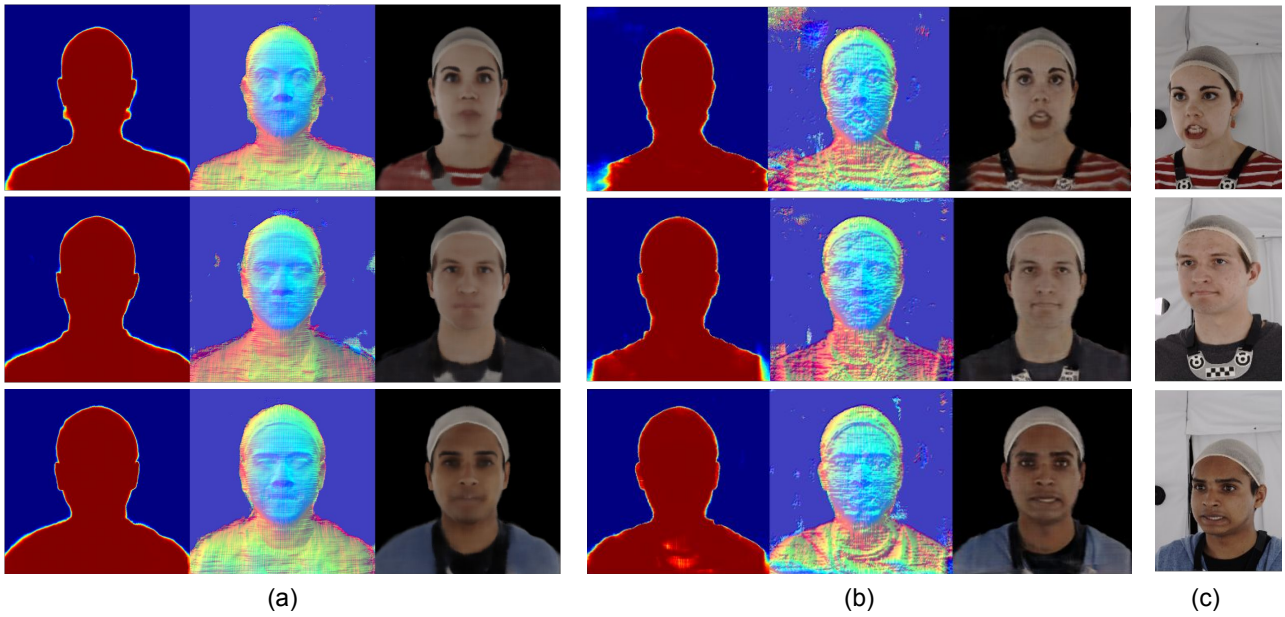

Figure 5: Generated alpha/normals/avatar in the canonical viewpoint using (a) eNerf and (b) Ours for (c) the ground truth identity. Note, for this experiment the eNeRF baseline has seen all identities and expressions at training time. Our approach not only better captures the identity of the person, but also the facial expression, while not having seen these specific identities at training time. We attribute this better generalization behaviour to our pixel-aligned features.

	SSIM (↑)	MSE(↓)	LPIPS (↓)
cNeRF	0.7663	1611.0112	4.3775
NV	0.8027	1208.36	3.1112
Ours	**0.8889**	**383.71**	**1.7392**

Table 1: Quantitative comparison of our approach (Ours) to reconstructions from, Neural Volumes (NV), and Globally conditioned NeRF (cNeRF).

of completeness and the amount of reconstructed details. Our method produces more complete reconstructions than Reality Capture, which would require many more views of the person to obtain a good reconstruction. In addition, our approach also leads to more detailed reconstruction than the globally conditioned Neural Volumes and cNerf approaches. We attribute this better generalization to the use of the pixel-aligned features, that better inform the model at test time.

4.4. Quantitative Comparisons

We compare the performance of our method with NV and cNeRF baselines (we omit RC because it fails to capture the complete head shape) in Table 1 on three common metrics from the literature (SSIM, LPIPS[32] and MSE). We note that our framework outperforms all the baselines by a considerable margin.

4.5. Analysis

We observe that methods that use global identity encoding like Neural Volumes and cNeRF do not generalize well to unseen identities as these methods are designed to be trained in a scene specific manner. Particularly, we notice in cNeRF that the facial features are smoothed out and some of the local details of unseen identities (like facial hair in row 3 and 4, and hair length in row 2) are lost, since this model relies heavily on the learned global prior. Reality Capture fails to capture the structure of the head as there are no priors built into the SfM+MVS framework, leading to incomplete reconstructions. A large number of images would be required to faithfully reconstruct a novel identity using RC (we refer to the supplementary document for additional analysis). Neural volumes is able to generate better textures because of the generated warp field which accounts for some degree of local information. However, since neural volume uses an encoder-decoder architecture, with the encoder using a global encoding, it projects test time identities into the nearest training time identity leading to inaccurate avatar predictions. Our proposed framework is able to reconstruct volumetric heads from just two example viewpoints, along with the structure of the hair.

Expression Information We present additional qualitative comparison on the ability of our model to better capture expression information in Fig. 5. We train another conditional NeRF baseline for expressions. Particularly, since cNeRF cannot generalize to novel identities, we train a NeRF model conditioned on a one-hot expression code and one-hot identity information (eNeRF) on test time identities (unseen for our method). We observe in this case that despite having seen all the identities during training eNeRF fails to generalize to dynamic expressions for multiple identities. Since our method leverages the local features for conditioning, it is better able to capture dynamic effects on a specific identity (both geometry and texture).

Num. Views	SSIM (↑)	MSE(↓)	LPIPS (↓)
1	0.8467	1467.15	2.9486
2	0.8596	1314.17	2.6451
3	0.8632	1285.67	2.5582
4	0.8739	1191.08	2.3606
5	**0.8753**	**1181.32**	**2.3167**

Table 2: Quantitative evaluation of the number of required conditioning views.

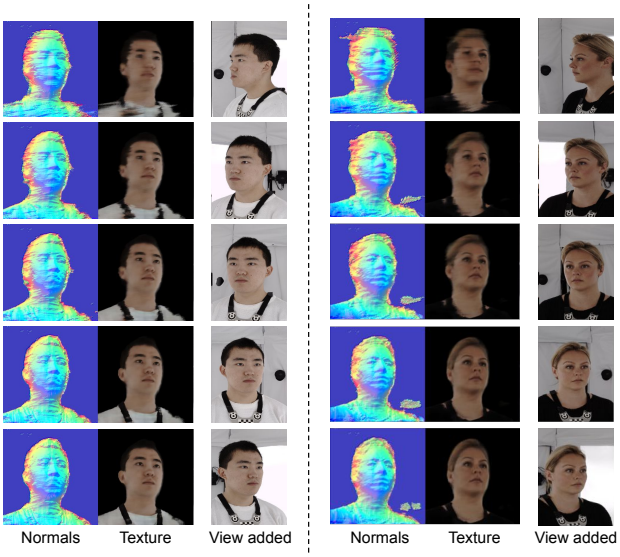

Figure 6: Predicted texture with respect to the number of views. In each row, we add one additional conditioning view (top to bottom). As can be seen, each added input increases the reconstruction quality.

5. Ablation Studies

In the following, we perform several ablation studies to explore different aspects of our approach in more detail.
How does the quality of the generated images change with respect to the number of example images? Fig. 6 shows view extrapolation for unseen identities. Particularly, since our model learns shape priors from training identities, the predicted normals are consistent with the input identity.

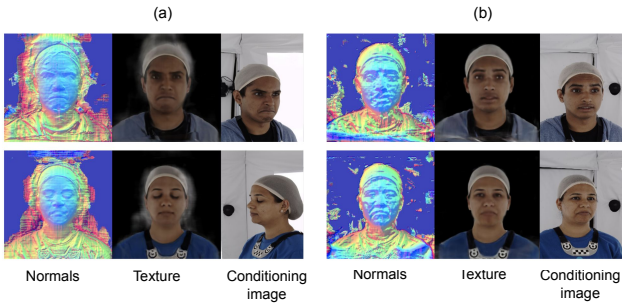

Figure 7: Background ablation. (a) Without background estimation (b) Ours. Our learned background model leads to better reconstruction results.

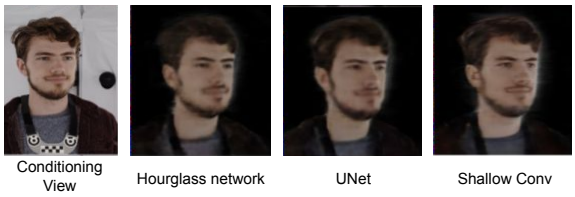

Figure 8: We demonstrate the sensitivity of the pixel-aligned features to the choice of the employed feature extractor. As can be seen, our shallow convolutional network leads to better reconstructions.

However, when extrapolating to extreme views (1st row), artifacts appear in the parts of the face that are unseen in the conditioning images. This is because of the inherent depth ambiguity due to projection of the sample points onto the feature image. We see that adding just the second view already significantly reduces these artifacts as the model now has more information regarding features from different views and can thus reason about depth. In practice, we find that we can achieve a large degree of view extrapolation with just two conditioning views. Tab. 2 provides the corresponding quantitative evaluation.

Is camera information required in addition to the extracted features? Fig. 9 demonstrates the need to incorporate camera information in the extracted features. Particularly, without the camera information, we see a large degree of streaking in the generated images due to inconsistent averaging of information from different viewpoints (particularly in row 1 and 2).

Are the results sensitive to the employed feature extraction network? U-Net and hour glass networks are some of the popular feature extraction networks used in recent works [23]. However, we find that in our setting a shallow encoder-decoder architecture serves as the best feature extraction networks (Fig. 8) as it preserves more of the local information without having to encode all the pixel level information into a bottleneck layer.

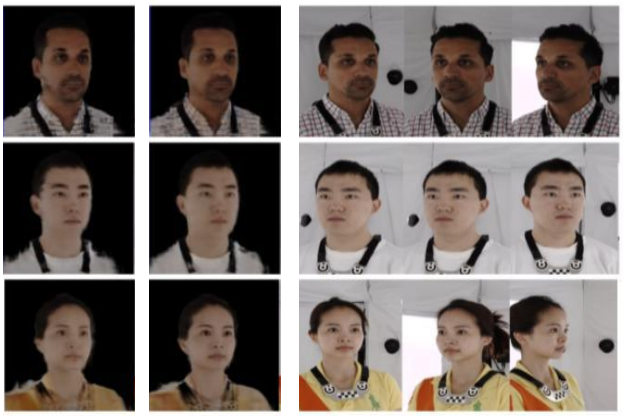

Figure 9: Feature summarization. As can be seen, our camera-aware feature summarization strategy leads to higher quality results than using simple mean pooling.

6. Limitations

While we have demonstrated compelling results for predicting volumetric avatars of human heads from just a small number of example images, our approach is still subject to a few limitations that can be addressed in follow-up work: (1) Our approach currently has limited extrapolation capabilities in terms of completely unobserved regions, e.g., the back of the head will not be reconstructed in detail if only front views are provided as example images. The incorporation of a global prior could improve generalization in such scenarios. (2) Our approach can currently not be applied to in-the-wild data. This has multiple reasons: First, we require the absolute head pose at test time for each of the example images. Second, our training corpus does not capture the spectrum of illumination and background variation of in-the-wild images. This could be tackled in the future by a more sophisticated training corpus or by data augmentation strategies.

7. Conclusion

We presented PVA - a novel approach for predicting volumetric avatars of the human head given only a small number of images as input. To this end, we devised a neural radiance field that leverages local, pixel-aligned features that can be extracted directly from the inputs, thus side-stepping the need for very deep or complex neural networks. Our approach is trained in an end-to-end manner solely based on a photometric re-rendering loss *without requiring* explicit 3D supervision. We have demonstrated that our approach outperforms the existing state of the art in terms of quality and that we are able to generate faithful facial expression in a multi-identity setting. We hope that this approach will serve as a simple and strong baseline for future work.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE international conference on computer vision*, pages 4432–4441, 2019. 2

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 2

[3] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In *Advances in Neural Information Processing Systems*, pages 9609–9619, 2019. 2

[4] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 484–498. Springer, 1998. 2

[5] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9984–9993, 2019. 3

[6] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 2

[7] 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, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019. 2

[8] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2018. 2

[9] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields, 2020. 1

[10] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[11] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7708–7717, 2019. 2

[12] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Trans. Graph.*, 37(4), July 2018. 2

[13] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4), July 2019. 1, 3, 4

[14] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. *arXiv preprint arXiv:2008.02268*, 2020. 3

[15] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 4

[16] 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. *arXiv preprint arXiv:2003.08934*, 2020. 1, 3, 4

[17] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. pagan: real-time avatars using dynamic textures. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018. 2

[18] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7588–7597, 2019. 3

[19] Thu H Nguyen-Phuoc, Chuan Li, Stephen Balaban, and Yongliang Yang. Rendernet: A deep convolutional network for differentiable rendering from 3d shapes. In *Advances in Neural Information Processing Systems*, pages 7891–7901, 2018. 2

[20] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. In *Proceedings IEEE International Conf. on Computer Vision (ICCV)*, 2019. 3

[21] Kyle Olszewski, Sergey Tulyakov, Oliver Woodford, Hao Li, and Linjie Luo. Transformable bottleneck networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7648–7657, 2019. 2

[22] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019. 2, 3, 4

[23] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 3, 8

[24] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[25] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of Computer Vision and Pattern Recognition (CVPR 2019)*, 2019. 2

[26] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 2

[27] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in*

Neural Information Processing Systems, pages 1121–1132, 2019. 2, 3

[28] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020. 2

[29] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. *arXiv preprint arXiv:2004.03805*, 2020. 3

[30] Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. Stylegan2 distillation for feed-forward image manipulation. *arXiv preprint arXiv:2003.03581*, 2020. 2

[31] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[33] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. In *Advances in neural information processing systems*, pages 118–129, 2018. 2