

Reconstructing Personalized Semantic Facial NeRF Models From Monocular Video

XUAN GAO*, University of Science and Technology of China, China
 CHENGLAI ZHONG, University of Science and Technology of China, China
 JUN XIANG, University of Science and Technology of China, China
 YANG HONG, University of Science and Technology of China, China
 YUDONG GUO, Image Derivative Inc, China
 JUYONG ZHANG†, University of Science and Technology of China, China

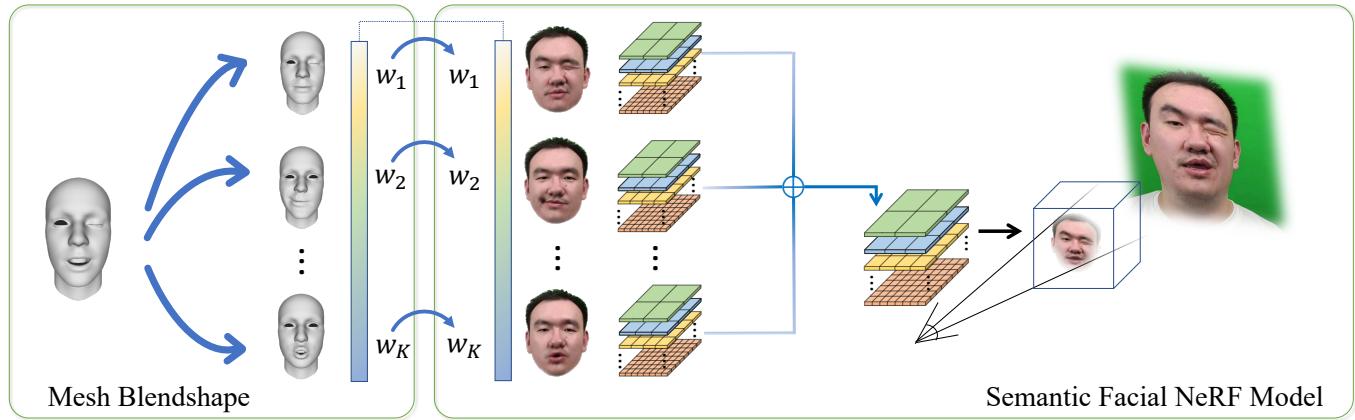


Fig. 1. A semantic model for human head defined with neural radiance field is presented. In this model, multi-level voxel field is adopted as basis with corresponding expression coefficients, which enables strong representation ability on the aspect of rendering and fast training.

We present a novel semantic model for human head defined with neural radiance field. The 3D-consistent head model consist of a set of disentangled and interpretable bases, and can be driven by low-dimensional expression coefficients. Thanks to the powerful representation ability of neural radiance field, the constructed model can represent complex facial attributes including hair, wearings, which can not be represented by traditional mesh blendshape. To construct the personalized semantic facial model, we propose to define the bases as several multi-level voxel fields. With a short monocular RGB video as input, our method can construct the subject's semantic

facial NeRF model with only ten to twenty minutes, and can render a photo-realistic human head image in tens of miliseconds with a given expression coefficient and view direction. With this novel representation, we apply it to many tasks like facial retargeting and expression editing. Experimental results demonstrate its strong representation ability and training/inference speed. Demo videos and released code are provided in our project page: <https://ustc3dv.github.io/NeRFBlendShape/>

CCS Concepts: • Computing methodologies → Reconstruction.

Additional Key Words and Phrases: Blendshape, Neural Radiance Field, Facial Retargeting

ACM Reference Format:

Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. 2022. Reconstructing Personalized Semantic Facial NeRF Models From Monocular Video. *ACM Trans. Graph.* 41, 6, Article 200 (December 2022), 12 pages. <https://doi.org/10.1145/3550454.3555501>

1 INTRODUCTION

3D face/head representation is an important research problem in computer vision and computer graphics, and has wide applications in AR/VR, digital games and movie industry. How to represent the dynamic head and faithfully reconstruct a personalized human head model from a monocular RGB video is an important and challenging research topic. With the hypothesis that human head could be embedded into a low dimensional space. Parametric semantic head models, like blendshape, have been studied and improved for a long

*This work was done when Xuan Gao, Chenglai Zhong and Jun Xiang were intern at Image Derivative Inc.

†Corresponding author (juyong@ustc.edu.cn).

Authors' addresses: {Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Juyong Zhang}, University of Science and Technology of China, 96 Jinzhai Road, Hefei 230026, Anhui, China, {gx2017@mail.ustc.edu.cn, zcl2017@mail.ustc.edu.cn, xiangjunxjk1@mail.ustc.edu.cn, hymath@mail.ustc.edu.cn, juyong@ustc.edu.cn}; Yudong Guo, Image Derivative Inc, 998 Wenyi West Road, Hangzhou, Zhejiang, China, guoyudong@idr.ai.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.
 0730-0301/2022/12-ART200 \$15.00
<https://doi.org/10.1145/3550454.3555501>

time. The blendshape head model, in the form of linear/bilinear combination of different facial expressions, has the following advantages. It is a semantic parameterization. The combination coefficients have intuitive meaning for the users as the strength or influence of specific facial expressions. Meanwhile, the blendshape constructs a reasonable shape space which can help the user freely control and edit in the space.

The generalized semantic head models like FaceWarehouse [Cao et al. 2013] aim to model different subjects with different expressions, and thus may ignore personalized geometry and texture details. To construct a personalized blendshape model, traditional mesh based methods usually adopt deformation transfer[Garrido et al. 2016; Li et al. 2010; Sumner and Popović 2004] and multilinear tensor-based 3DMM[Cao et al. 2018, 2014; Vlasic et al. 2006]. However these methods usually have the following disadvantages. First, mesh based parametric models are hard to represent personalized non-face parts like hair and teeth. Second, to use an RGB supervision, we have to use approximated differentiable rendering techniques to alleviate the non-differentiable problems. Third, deformation transfer cannot reconstruct expressions realistically due to limited representation ability. Last, facial expressions are characterized by many factors such as ages and muscle movements, and these factors are hard to be accurately expressed by predefined blendshapes.

Recently, NeRF based methods have made it possible to synthesize photorealistic images. Some works integrate NeRF with GANs[Chan et al. 2022, 2021; Deng et al. 2022; Gu et al. 2022; Niemeyer and Geiger 2021; Schwarz et al. 2020; Zhou et al. 2021]. However, this kind of generative models couple expression, identity and appearance together, resulting that the expressions can not be easily controlled. HeadNeRF[Hong et al. 2022] proposes to disentangle different semantic attributes, but it could not represent personalized facial dynamics and facial details due to its generic model capacity. AD-NeRF[Guo et al. 2021b] and NerFACE[Gafni et al. 2021] could generate highly personalized facial animation, their user-specific training make the model learn more personalized facial details. However, they need a long time to train a reasonable dynamic head field. According to our experiments in section 4.3, this is because they concatenate the expression condition with Fourier positional information and directly input it to the MLP. Both the Fourier positional encoding and the “concatenate” strategy is not ideal for fast training. The Fourier encoding is not friendly for MLP for fast convergence. And the concatenation operation does not contain any combination law to discover the relation between local and global features (in NeRF case, positional information and expression condition). Therefore, it takes a long time for MLP to learn how to use the expression condition to predict RGB and density.

Recently, local features have been explored to improve NeRF’s quality and efficiency. The original NeRF’s local feature is the Fourier positional encoding, which takes a long time to converge. Following works designs different kinds of local features to improve NeRF. Some methods adopt a voxel field to accelerate the training process[Sara Fridovich-Keil and Alex Yu et al. 2022; Sun et al. 2022]. Other works use the voxel field to accelerate ray marching and volume rendering[Garbin et al. 2021; Liu et al. 2020; Yu et al. 2021a]. EG3D[Chan et al. 2022] adopts a compact and efficient tri-plane architecture enabling geometry-aware synthesis. TensoRF[Chen et al.

2022] factorizes the 4D scene tensor into multiple compact low-rank tensor components to separate local features. Among these methods, instant neural graphics primitives (INGP)[Müller et al. 2022] demonstrated a remarkable performance improvement in both training and rendering. It uses a highly compressed compact data structure, multi-level hash table, to make it possible to store a multi-level voxel field. A novel design of INGP is that the feature query collision is solved in an adaptive way. Features in different levels could be trained together. Together with the help of high-performance ray-marching implementation, it could train a static NeRF scene less than 1 min and render one frame in tens of milliseconds.

Although a lot of methods have been proposed to speed up the training and inference of a static NeRF field, it still remains a problem to achieve a fast training of a dynamic scene such as the complicated head deformation. As our baseline shows, using a direct “concatenate” strategy to combine local features and expression code together as the input of MLP, which is very common for NeRF based head application, is not efficient and sufficient to model dynamic head motions.

In this paper, we present a personalized semantic facial model architecture defined on multi-level voxel field. It not only inherits the semantic meaning of mesh blendshape used for tracking, but also has more personalized facial detailed attributes especially for no-face part. Each basis of our model is a radiance field of a specific expression, represented by a multi-level voxel field. We adopt the multi-resolution hash tables to store the multi-level features for performance consideration. For any novel expressions, it can be expressed as the weighted combination of voxel bases with the expression coefficients. We adopt an MLP to interpret the voxel field as a radiance field for volume rendering. To further accelerate the ray marching in volume rendering and make the optimization focus on the region possibly occupied by head, we design an expression-aware density grid update strategy. Thanks to powerful representation ability and fast convergence of our implicit model, our method outperforms other similar head construction methods in both modeling quality and construction speed. Our method can construct a photo-realistic personalized semantic facial model in around 10-20 minutes, which is remarkably faster than related NeRF based head technique. As our model is trained from a video of a specific person and combines the features in a latent space, it could capture personalized details including non-linear deformation (cheek folds, forehead wrinkle) and user-specific attributes(mole, beard). Compared with traditional mesh based blendshape models, our model can be constructed from a short RGB video and generate high-fidelity view consistent head images with different expressions.

In summary, the contributions include the following aspects:

- We present a novel semantic model for human head defined with neural radiance field. Our constructed NeRF basis not only has a disentangled semantic meaning, but also embodies more personalized facial attributes including muscle actions and detailed texture. Therefore, the constructed digital avatars can model facial motions quite well and generate photo-realistic results.
- Our representation combines multi-level voxel fields with expression coefficients in the latent space. The multi-resolution

features could efficiently learn head details in different scales. The linear blending design could modulate the local features in advance to adapt to MLP's input distribution, which makes our model cost much less time to construct and express more realistic facial details.

- With this novel representation, digital human head related applications like facial reenactment can be easily achieved and have remarkable performance, which implies its potential usage in photo-realistic animation industry.

2 RELATED WORK

Parametric Head Model. Under the hypothesis that human head shape space can be well disentangled as identity, expression and appearance, Blanz and Vetter proposed 3DMM[Blanz and Vetter 1999] to embed 3D head shape into several low-dimensional PCA spaces. Mesh based parametric head model has been further studied by a lot of following works. To improve its representation ability, some work extends it to multilinear models[Cao et al. 2013; Vlasic et al. 2006], non-linear models[Guo et al. 2021a; Ranjan et al. 2018; Tran and Liu 2018] and articulated models with corrective blendshape to improve its modeling ability[Li et al. 2017]. Both mesh based methods and deep learning based methods have been widely used in many related applications. However, mesh based parametric models usually can not represent personalized facial details due to its limited representation ability. Meanwhile, existing mesh based parametric models can not represent non-face parts especially for hair. Some works handle this problem using deformation transfer[Cao et al. 2016; Garrido et al. 2016; Hu et al. 2017; Ichim et al. 2015; Sumner and Popović 2004] or neural network[Bai et al. 2021; Chaudhuri et al. 2020; Yang et al. 2020] to get user-specific blendshape basis.

To break through the limited representation ability of explicit mesh based digital human representation, many works adopt the implicit representation to improve the model capacity and visual quality [Gafni et al. 2021; Hong et al. 2022; Jiang et al. 2022; Wang et al. 2022; Yenamandra et al. 2021; Zheng et al. 2022; Zhuang et al. 2021]. i3DMM[Yenamandra et al. 2021] is the first neural implicit function based 3D morphable model of full heads. HeadNeRF[Hong et al. 2022] proposes a generic head parametric model based on neural radiance field. Although neural implicit function based representations have demonstrated strong representation ability, a generic model often still lacks personalized facial details. NerFACE[Gafni et al. 2021] presents a personalized NeRF based human head model. However, their method requires a long time for training and inference for each subject. IM Avatar[Zheng et al. 2022] presents an implicit LBS model. Note that both our method and IM Avatar have an implicit blendshape architecture. The main difference is that IM Avatar focuses on detailed geometry and appearance and our model focuses more on photorealistic rendering and efficient training/inference. Another difference is that IM Avatar uses a backward non-rigid ray marching to find the canonical surface point for each ray. Our ray marching is performed in the deformed space.

Human Portrait Synthesis. Many methods have been proposed for facial reenactment and novel view synthesis. Image based methods[Pumarola et al. 2019; Siarohin et al. 2019; Zakharov et al. 2020]

adopt warping fields or encoder-decoder architectures to synthesize the images. As these methods represent the 3D deformation in the 2D space, artifacts may appear for large pose and expression changes. Morphable model[Kim et al. 2018; Thies et al. 2020a, 2019, 2016; Zhang et al. 2022] based methods use a parametric 3D model to synthesize a digital portrait. Deep Video Portraits[Kim et al. 2018] uses rendered correspondence maps together with an image-to-image translation network to output photo-realistic imagery. Deferred Neural Rendering[Thies et al. 2020a, 2019] proposes an object-specific neural textures which can be interpreted by a neural renderer.

Neural Radiance field. NeRF[Mildenhall et al. 2020] proposes to represent a scene with an MLP and utilize the volume rendering for novel view synthesis task. As NeRF is differentiable, its inputs can be only multi-view images. Due to the above listed characteristics, NeRF has been widely used to 3D geometry reconstruction[Wang et al. 2021; Yariv et al. 2021], 4D scene synthesis [Li et al. 2022; Park et al. 2021a,b] and digital human modeling[Peng et al. 2021a,b; Weng et al. 2022],etc. Besides, a lot of research focus on improving NeRF's representation ability[Barron et al. 2021] and reducing the number of inputs[Chibane et al. 2021; Niemeyer et al. 2022; Yu et al. 2021b].

Recently, NeRF has also demonstrated its strong representation ability in human head modeling. Many works adopt NeRF to represent dynamic human head scene, and synthesis high-fidelity 3D consistent result. Generative head models [Chan et al. 2022, 2021; Deng et al. 2022; Gu et al. 2022; Niemeyer and Geiger 2021; Schwarz et al. 2020; Zhou et al. 2021] use latent code to generate the rendering result. Although they usually have a good pose control over the result, but do not support expression editing due to its generative adversarial training strategy. Generic parametric head model [Hong et al. 2022; Zhuang et al. 2021] disentangles latent space of human head as identity, expression and appearance space, and to some extent realize semantical control over head transformation. However, generic head model often ignores personalized facial details and user-specific facial muscle movements due to limited MLP capacity. AD-NeRF[Guo et al. 2021b] and NerFACE[Gafni et al. 2021] are subject-specific models, and it can generate high fidelity human head animation controled by voices or expressions. However, both AD-NeRF and NerFACE need days for training and seconds for inference. And we found both of them tend to learn a smooth head scene and sometimes ignore high-frequency facial attributes.

Voxel Representation for NeRF Acceleration. With the help of voxel field, NeRF could spare its training burden across the local features, which will significantly improve the training speed[Sara Fridovich-Keil and Alex Yu et al. 2022; Sun et al. 2022]. Voxel field could also help store the spacial information like density distribution in advance to accelerate the inference speed[Garbin et al. 2021; Liu et al. 2020; Lombardi et al. 2021; Yu et al. 2021a]. Recently, instant neural graphics primitives[Müller et al. 2022] adopts multi-level hash table to augment a shallow MLP and achieves a combined speedup of several orders of magnitude. It can train a static scene with NeRF using only several seconds, and render the scene in tens of milliseconds. However, these methods could not be directly used for dynamic scenes due to its complex non-rigid deformations. Meanwhile, it is hard to perform "pruning" operation for voxel grid

in dynamic scenes, which is usually important for ray-marching and information storage.

3 METHOD

In this work, we propose a novel personalized human head representation that takes a series of specially designed neural radiance fields as the bases of the human head. Similar with traditional mesh blendshape like FaceWarehouse [Cao et al. 2013], each basis of the proposed model has a specified semantic meaning, such as eye closed and jaw forward, which makes it easy for users to use a low-dimensional code to generate desired human head images.

Our head model linearly combines multi-level voxel fields with expression coefficients in the latent space, and the multi-resolution features could efficiently learn head details in different scales. Our linear blending design could modulate the local features in advance to adapt to MLP’s input distribution, which reduces the training burden of MLP and expresses more realistic facial details. We improve the rendering efficiency of our model by concentrating the sampling near surfaces. With these designs, the proposed method could construct a set of NeRF bases in less than 20 minutes. Meanwhile, the trained model has interactive rendering speed and can render photo-realistic human head images. Furthermore, the results generated by our personalized NeRF-based blendshape can be further semantically edited, such as freely adjusting camera parameters and independently changing the subject’s expression to any desired expression while keeping other attributes unchanged. An overview of the proposed representation is shown in Fig. 2, and the algorithm details will be presented in the following.

3.1 NeRF based Linear Blending Representation

Similar with [Chan et al. 2022; Gafni et al. 2021; Hong et al. 2022], our representation is also based on neural radiance field [Mildenhall et al. 2020], and we represent it by the MLP-based implicit function. Besides, we associate expression bases with multi-level hash tables [Müller et al. 2022] and endow each hash table with specified semantic attributes via an elaborately designed training strategy. We denote the representation of our model as \mathcal{R} and formulate it as:

$$I = \mathcal{R}_\theta(C, \mathbf{h}_0 + \mathbf{Hw}), \quad (1)$$

where θ indicates the learnable weight parameters of the MLP. C is the camera parameter used for rendering, including the extrinsic and intrinsic matrices. $\mathbf{h}_0 \in \mathbb{R}^{L \times T \times F}$ is the multi-level hash table representing the mean shape of the blendshape in latent space, where L is the number of hash table’s levels, T is the hash table size, and F is the number of feature dimensions per entry of the hash table. $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_K\}, \mathbf{h}_i \in \mathbb{R}^{L \times T \times F}$ is the multi-level hash table representing the expression displacement basis in latent space. K is the number of the expression bases of our model, and $\mathbf{w} = \{w_1, w_2, \dots, w_K\} \in \mathbb{R}^K$ is the expression coefficient. I is the human head image rendered by \mathcal{R}_θ according to the above parameters.

3.2 Rendering

The rendering process of our model is shown in the right part of Fig. 2. We first calculate the corresponding hash tables for a given

expression coefficient \mathbf{w} as:

$$\mathbf{h} = \mathbf{h}_0 + \mathbf{Hw} = \mathbf{h}_0 + \sum_{i=1}^K w_i \mathbf{h}_i, \mathbf{h} \in \mathbb{R}^{L \times T \times F}, \quad (2)$$

where we combine the bases in multi-level voxel space instead of combining the blendshape basis in explicit space as mesh blendshape. Then we adopt the model architecture of [Müller et al. 2022] and formulate the MLP-based implicit function g_θ of NeRF as:

$$g_\theta : (\eta(\mathbf{x}; \mathbf{h}), \gamma(\mathbf{d})) \mapsto (\sigma, c), \quad (3)$$

where $\mathbf{x} \in \mathbb{R}^3$ is a 3D point sampled from one ray. $\mathbf{d} \in \mathbb{R}^3$ indicates a unit vector representing view direction. $\eta(\mathbf{x}; \mathbf{h}) \in \mathbb{R}^{L \times F}$ is the encoding of \mathbf{x} about \mathbf{h} , and it is obtained by linearly interpolating the feature vectors indexed by the hash value of \mathbf{x} ’s transformed integer corner points (please refer to [Müller et al. 2022] for details). $\gamma(\mathbf{d})$ is the positional encoding of \mathbf{d} , which projects \mathbf{d} onto the first 16 coefficients of the spherical harmonics basis. σ and c denote the predicted density and color of \mathbf{x} , respectively.

Actually, the expression coefficients could combine the local features encoded by the multi-level hash tables, and this effectively enhances the representation ability of our model. Thus, we can use a lightweight MLP to represent the implicit function g_θ . The network architecture of g_θ is a 4 layers MLP with 64 neurons width. It means that the rendering of our model can be executed efficiently. Lastly, we generate the rendered image I by the following volume rendering:

$$I(r) = \int_0^\infty p(t)c(r(t))dt, \quad (4)$$

where $p(t) = \exp(-\int_0^t \sigma(r(s))ds)\sigma(r(t)).$

$r(t)$ represents a ray emitted from the camera origin. The head mask can be generated in a similar way:

$$M(r) = \int_0^\infty p(t)dt. \quad (5)$$

In summary, the rendering process of our model consists of the following steps: First, we use the expression coefficients to linearly combine the multi-level hash tables that represent different expression bases in latent space. Then we cast rays to get sampled points. By querying these points in the combined hash table, the hash encoding of the sample point w.r.t the combined hash table is generated. We further use a lightweight MLP-based implicit function to map the generated hash encoding to RGB and density for volume rendering. We want to point out that these steps can be fast calculated, and we will further skip empty space when building our model. Therefore, the rendering of our model is efficient, and the constructed personalized semantic head model can quickly generate the target head image with the given target expression coefficient. Besides, efficient rendering actually speeds up the construction of our models as well.

3.3 Construction

The proposed personalized NeRF-based semantic facial model can be constructed using only a 3-5 minutes monocular RGB video thanks

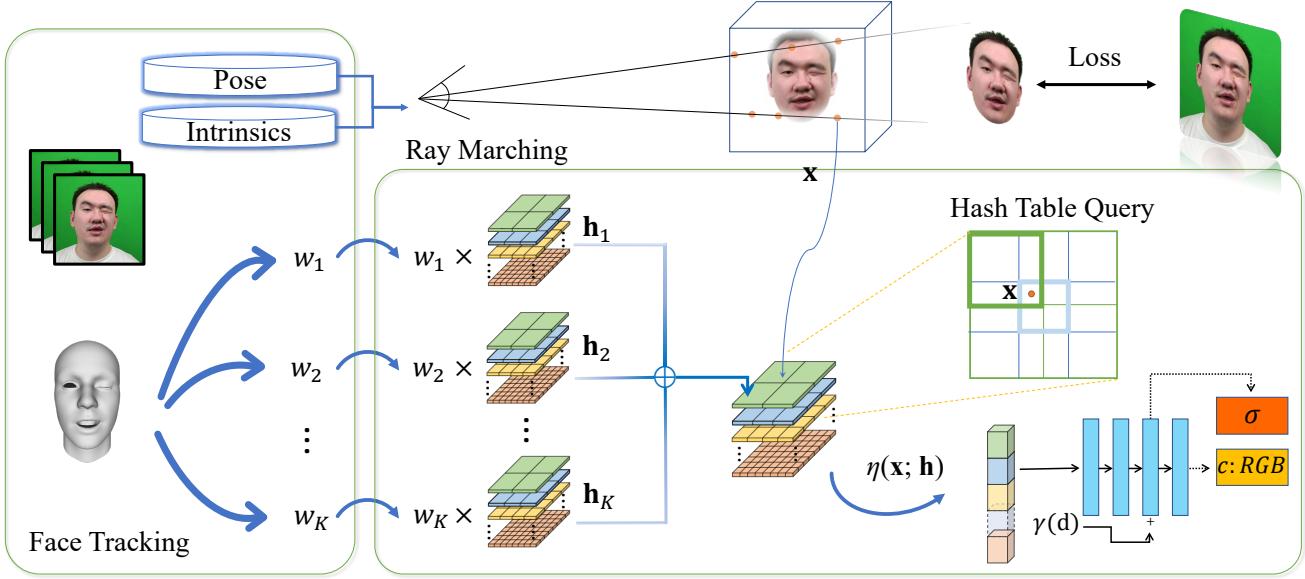


Fig. 2. Our pipeline, we track the RGB sequence and get expression coefficients, poses and intrinsics. Then we use the tracked expression coefficients to combine multiple multi-level hash tables to get a hash table corresponding to a specific expression. Then the sampled point is queried in hash table to get voxel features, we use an MLP to interpret the voxel features as RGB and density. We fix the expression coefficients and optimize the hash tables and MLP to get our head model.

to the above-mentioned elaborate design. In the following, we will present the algorithm details to construct our model.

3.3.1 Data Preprocessing. Firstly, we use an existing mesh-based facial blendshape [Cao et al. 2013] to track the face in the input video similar to [Guo et al. 2018]. Then we can obtain the expression coefficients and the head pose parameters of each frame. Like Head-NeRF [Hong et al. 2022], we take the human head pose parameters as the extrinsic camera parameter of the corresponding frame, which implicitly aligns the underlying geometry of each frame to the same spatial location. Lastly, we randomly extract N frames from the input video to train our model, and the head mask of the selected frame is generated by the off-the-shelf segmentation methods[Yu et al. 2018]. We denote the optimized expression coefficients of selected frames as $\{w^1, w^2, \dots, w^N\}$.

3.3.2 Training. The learnable variables of our model include the network parameters θ of the implicit function g_θ and the multi-level hash tables (\mathbf{h}_0, \mathbf{H}) representing expression bases. Meanwhile, we take $\{w^1, w^2, \dots, w^N\}$ as the expression coefficients of our personalized NeRF-based blendshape and freeze them while training our model. The loss terms used to train our model include the following three terms:

Photometric Loss. This loss requires that the rendered result is consistent with the input RGB image, which is common for NeRF-based reconstruction and can be formulated as:

$$L_{color} = \sum_{r \in \mathcal{S}} \|I(r) - I_{GT}(r)\|_1, \quad (6)$$

where \mathcal{S} is the set of rays in each batch, and $I(r), I_{GT}(r)$ are the predicted RGB colors and ground truth for ray r respectively.

Mask Loss. This loss requires that the rendered mask of Eq. (5) is consistent with the ground truth head mask. It is formulated as:

$$L_{mask} = \sum_{r \in \mathcal{S}} \|M(r) - M_{GT}(r)\|_1, \quad (7)$$

where $M(r)$ and $M_{GT}(r)$ are the predicted mask value and ground truth for ray r respectively. This loss makes sure density outside the head region is zero, and also lets the head region become an opaque object quickly.

Perceptual Loss. The perceptual loss L_{LPIPS} of [Zhang et al. 2018a] is utilized to provide robustness to slight misalignments and shading variations and improve details in the reconstruction. We choose VGG as the backbone of LPIPS. As LPIPS uses convolutions to extract features, we sample B patches with size $W \times W$, and render a total of $B \times W \times W$ rays in each batch (B is also the number of frames in each batch). The rendered patch is compared against the patch with the same position on the input image. Similar strategy is used in [Schwarz et al. 2020; Weng et al. 2022].

In summary, the overall loss of training our model is defined as:

$$L_{total} = \lambda_1 L_{color} + \lambda_2 L_{mask} + \lambda_3 L_{LPIPS}, \quad (8)$$

where λ_i is a scalar for balancing different terms. To minimize the total loss, we propose a well designed training strategy, and it contains three steps:

In the first two epochs, we set λ_1, λ_2 to 1 and λ_3 to 0. The mask loss could make the model quickly learn the distribution of 3D density because it is a supervision directly on the density distribution.

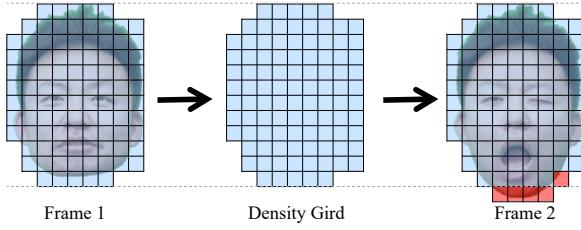


Fig. 3. The density grid of a specific expression may not cover all the expression cases. Heads in some frames may be out of the range.

Stable density distribution is also good for the following density grid update.

Although the mask loss could help the convergency of density field, we found the parsing mask is not very accurate especially for the hair part. Therefore, for 2nd-7th epoch, we set λ_2, λ_3 to zero and use photometric loss only. We use RGB data to supervise the NeRF training to learn fine detailed color and geometry.

Lastly, we not only randomly sample rays, but also sample patches. For randomly sampled rays, we set λ_1 to 1 and λ_2, λ_3 to 0, for sampled patches, we set λ_1 to 0.1 and λ_3 to 0.1. We sample patches in mouth part in 1/2 probability and sample patches across the whole image otherwise.

3.3.3 Expression-Aware Density Grid Update. We use a 128^3 density grid to store local density information to instruct ray-marching to skip empty space. Different from the static scene considered in [Müller et al. 2022], the captured dynamic human head is changing over time with different poses and expressions. As the example shown in Fig. 3, the density field of a specific expression could not cover all expression cases. We should not use the density field of a certain expression to determine the density grid. In our implementation, we compute density grid of each basis as:

$$\hat{\mathbf{h}}_i = \mathbf{h}_0 + \hat{w}_i \mathbf{h}_i, \quad (9)$$

where $\hat{w}_i = \max_{j \in [1, N]} w_i^j$.

w_i^j is the i -th element of w^j . We compute the element-wise maximum of the density grids of all $\hat{\mathbf{h}}_i$ to get the final density grid to approximate the natural 3D range of head expression.

3.3.4 Discussion on Fast Training. Our model can converge quickly due to the following reasons:

Firstly, the relationship between global condition and local features (in our case, expression coefficients and queried positional features) have been considered in our architecture. This linear blending architecture could reduce the burden of MLP in learning how to use expression information to transfrom the positional information. We also validate our linear blending is much better than a direct "concatenate" strategy. See 4.3 for experimental details.

Secondly, our model could capture multi-scale details at the same time. Features in different levels of multi-resolution hash table could be jointly optimized.

Thirdly, ray marching steps in empty space are skipped. we use a density grid to conduct efficient ray sampling. The density grid

Table 1. The parameters of hash table

| Parameter | Value |
|--|------------------------|
| Number of levels | 16 |
| Hash table size | 2^{14} |
| Number of feature dimensions per entry | 4 |
| Coarsest resolution | 16 |
| Finest resolution | 1024 |
| Initial distribution | $U(-10^{-4}, 10^{-4})$ |

update considers every expression action to make sure every voxel occupied by the head region are considered.

4 EXPERIMENTS

4.1 Implementation Details

We implement our semantic facial model with Pytorch [Paszke et al. 2019], and the CUDA extension of Pytorch is employed to implement the raymarching and volume rendering parts. Our model is trained with Adam solver [Kingma and Ba 2014] with batch size 4. The dimension of our expression coefficients is 46, and we set the patch size of the perceptual loss to 32. The parameters of our hash table are listed in Tab. 1. All the results are tested on one RTX 3090 card.

Some training videos are from the datasets collected by Neural Voice Puppetry[Thies et al. 2020b] and SSP-NeRF[Liu et al. 2022]. Since the expressions in these videos are mainly normal talking styles which are not very challenging, we didn't do evaluation on the whole dataset. We capture some monocular videos with exaggerated expressions and large head rotations. In each captured video, the subject is asked to perform arbitrary expressions, and we have got the permissions from these subjects for research purposes. For each target person, we collect a 3-5 min monocular RGB video with 512×512 resolution and 30 fps. Thus we captured 5000-10000 frame images for a single person. The last 500 frames serve as the testing dataset, and we randomly extract 3000-4000 frames from the rest as our training dataset.

4.2 Comparison

In this part, we compare our model's rendering quality and representation ability with existing state-of-the-art facial reenactment or head modeling methods. Specifically, FOMM [Siarohin et al. 2019] uses a reference image and a driving video as inputs to generate a motion video sequence. NerFACE [Gafni et al. 2021] and Neural Head Avatars (denoted as NHA) [Grassal et al. 2022] use the same training data as ours. NerFACE is NeRF-based head modeling, which takes the concatenation of the expression code and positional encoding information as input. Neural Head Avatars explicitly reconstruct the full head on a FLAME model.

Fig. 4 shows the qualitative comparison between our model and the above methods. It can be observed that our model is superior to others. We found the 3D consistency of FOMM is not as good as others. This drawback may originate from its 2D CNN architecture without enough 3D space knowledge. NerFACE tends to learn low-frequency signals and struggles to model head detailed information. Neural Head Avatars have more personalized details than NerFACE.

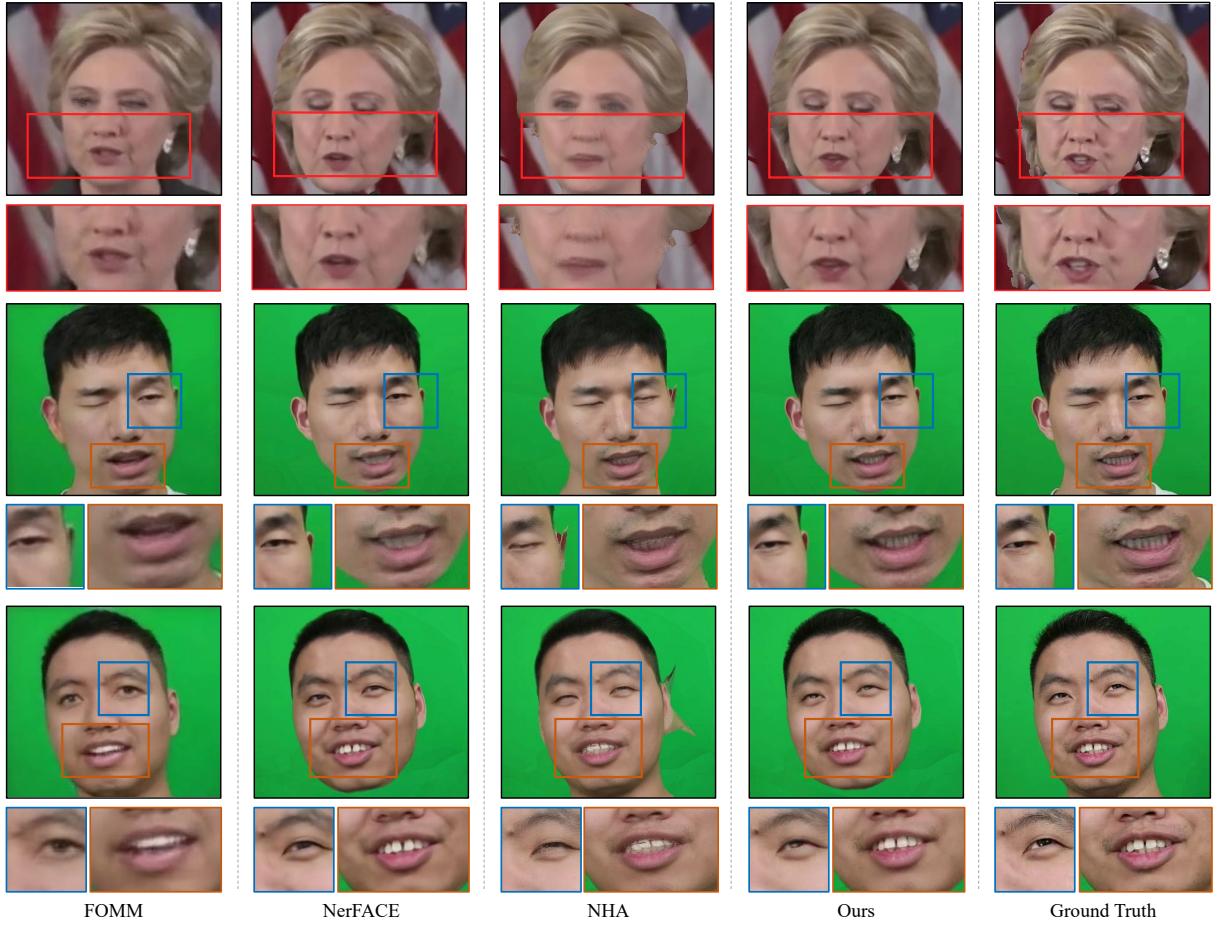


Fig. 4. Comparison with state-of-the-art head modeling and facial reenactment methods. We can see that our model reconstruct high-fidelity expressions and facial details. YouTube ID of Hillary Clinton's video is -yHgE9W699w.

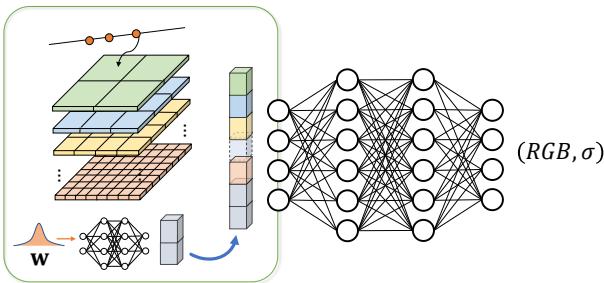


Fig. 5. Baseline architecture. The queried feature in hash table is concatenated with expression code as the input of MLP. We use a deeper and wider MLP to demonstrate its representation ability. A 2-layer MLP is used to map the expression coefficients to be concatenated with the queried feature.

However, the explicit mesh domain restricts its representation ability. As shown in this figure, undesired geometric artifacts appear in NHA's results. In contrast, our results are most consistent with the

ground truth, and the fine-level details of the human head can be represented and rendered.

Tab. 2 shows the quantitative comparison between our model and other methods, where the Mean Squared Error (MSE), L1 distance, Peak Signal-to-Noise Ratio (PSNR), Structure Similarity Index (SSIM) [Wang et al. 2004] and Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al. 2018b] are computed. It is worth noting that our method outperforms these existing methods in terms of any reconstruction error, which further verifies the effectiveness and superiority of our model.

As the expression coefficients space used for tracking is disentangled from identity space, we could use anyone's facial expression coefficients of the mesh-based blendshape to combine our bases and render the radiance field with a given view. Based on this observation, we apply our model to perform facial reenactment where we transfer the facial expressions from one person to another. Specifically, we track the source subjects' video and get poses, camera intrinsics, and expression coefficients. The poses and intrinsics are utilized to cast rays, and we further use expression coefficients to

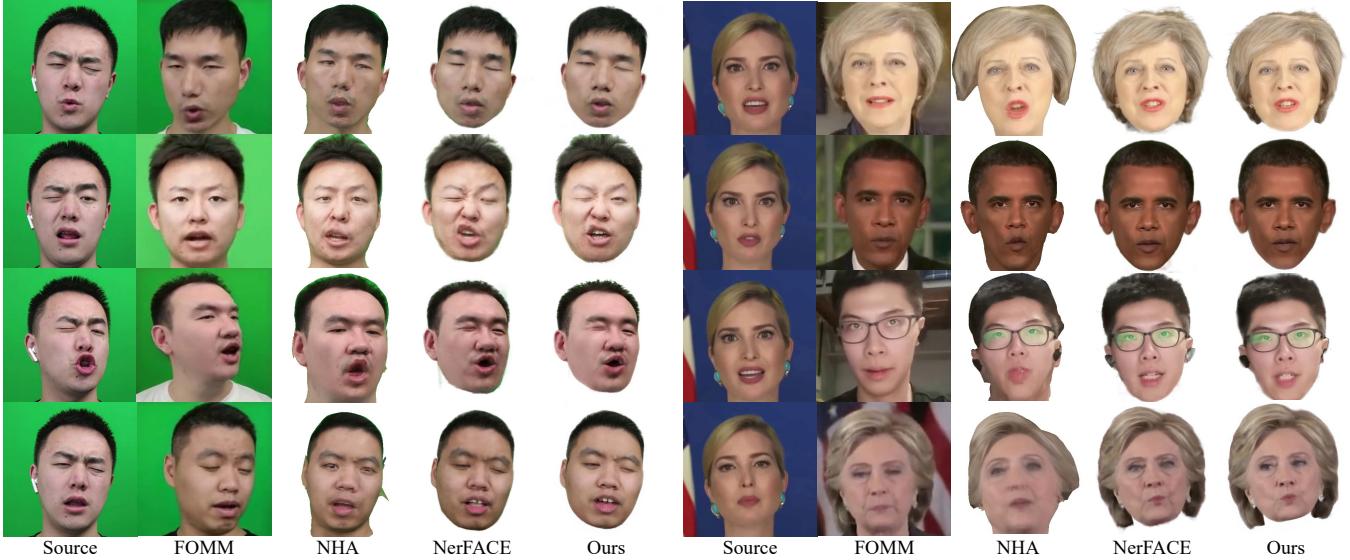


Fig. 6. Our model can be easily used for facial reenactment. Compared with other methods, our method also demonstrates more personalized facial details and synthesizes a more reasonable human head conditioned on expression coefficients in cross-identity reenactment. YouTube ID of Ivanka Trump’s video is -dNU9e0SYlg. YouTube ID of Theresa May’s video is nOj49CzODEU. YouTube ID of Barack Obama’s video is IQJW4_FvVKo. YouTube ID of Hillary Clinton’s video is -yHgE9W699w.

Table 2. Quantitative evaluation of our method in comparison to state-of-the-art facial reenactment methods based on self-reenactment. We compute the mean value and standard deviation of every method.

| Metrics | FOMM | NerFACE | NHA | Ours |
|---------------------------------|-------------|-------------|-------------|---------------------|
| MSE(10^{-3}) \downarrow | 1.75(0.81) | 0.75(0.45) | 0.69(0.54) | 0.48 (0.32) |
| L1(10^{-2}) \downarrow | 1.87(0.52) | 0.84(0.30) | 0.80(0.29) | 0.70 (0.23) |
| PSNR \uparrow | 28.32(2.45) | 32.24(2.70) | 32.85(3.01) | 34.15 (2.58) |
| SSIM(10^{-1}) \uparrow | 9.29(0.28) | 9.67(0.15) | 9.69(0.16) | 9.73 (0.13) |
| LPIPS(10^{-2}) \downarrow | 5.30(1.73) | 3.50(1.64) | 3.37(1.88) | 2.67 (1.32) |

combine targets’ bases. Then the sampled point is queried in the final hash table and interpreted by MLP to predict RGB and density. Compared with other methods, our method also demonstrates more personalized facial details and synthesizes a more reasonable human head conditioned on expression coefficients in cross-identity reenactment (See Fig. 6), which demonstrates that our bases are semantically correct and suitable for reenactment.

4.3 Comparison with “Concatenate” Operation

We want to point out that the multi-level voxel field is not the only reason for our model’s representation ability and training efficiency. In fact, the implicit linear blending architecture also plays an important role in the learning process of our model. To verify this, we design a “concatenate” baseline model. This baseline use only one hash table, and the queried feature is concatenated with expression coefficients as the input of the MLP-based implicit function. Note that the “concatenate” operation is frequently used in many NeRF-based head models[Gafni et al. 2021; Guo et al. 2021b; Hong et al. 2022]. In addition, We keep the training strategy the same and

use a much deeper MLP (7 layers to predict density and 1 layer to predict RGB, width 128) for this baseline to predict the RGB and density of sampled points. Fig. 5 shows the architecture of this baseline. Besides, we note that NerFACE [Gafni et al. 2021] is also a “concatenate” model, where the concatenation of the expression coefficients and the Fourier positional encoding is taken as the input of the MLP-base implicit function. Thus we regard NerFACE as another baseline of our method.

As shown in Fig. 7, our model could learn a dynamic head scene in less than 20 minutes, in which time neither the baseline nor NerFACE could get any plausible result. Meanwhile, we find that both baseline and our method could quickly learn the rigid part of the human head. The difference is that the baseline tends to learn the rigid region first and then the dynamic region, such as the eyes and mouth. In contrast, our model could learn much more dynamic relationships in a limited short time. As shown in this figure, our model could faithfully reconstruct facial expressions like the open mouth and closed eyes after training for only 5 minutes. The comparison implies that a simple “concatenated” input is hard for MLP to learn the facial dynamic region no matter whether a voxel feature is used. The reason of our superiority may be that our implicit blendshape architecture enables local features to be modified by expression coefficients in a global manner, which makes the local features adapt to MLP’s input distribution and reduces the learning burden of our model’s MLP. Fig. 8 depicts the relationship between PSNR and training time. We can see that our model’s training efficiency significantly outperforms both baseline and NerFACE.

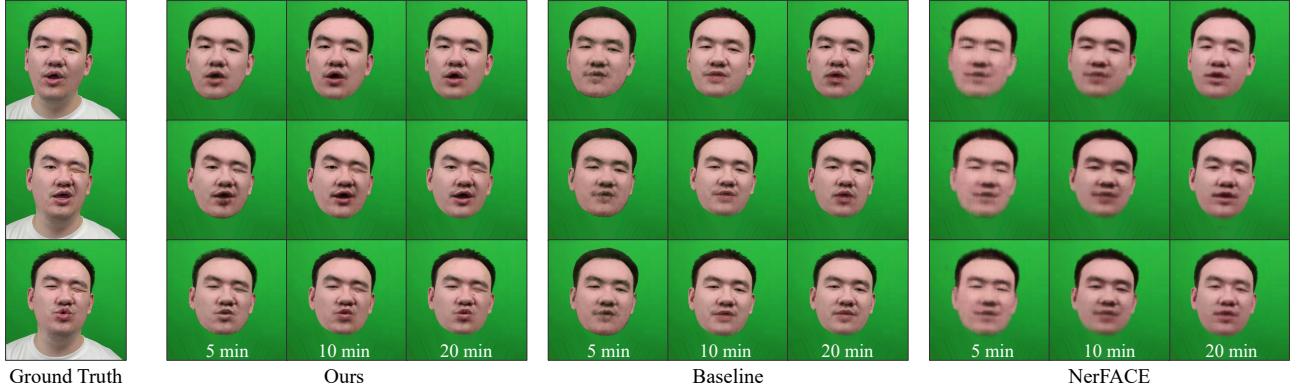


Fig. 7. Comparison with directly “concatenate” strategies including our baseline implementation and NerFACE [Gafni et al. 2021].

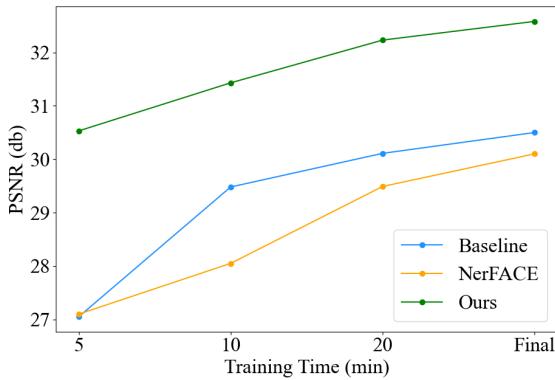


Fig. 8. Comparison with baseline and NerFACE on training speed.

4.4 Geometry Visualization

We extract the iso-surface from the density field with marching cubes. Although our model focuses more on photorealistic rendering and efficient training/inference, as shown in Fig. 10, we found that the geometry is reasonable. Note that we don't use any direct supervision on the geometry like normals or depths, but our model could still learn the geometric attributes like eyes, nose and detailed hair thanks to the radiance field representation. This also implies the 3D consistency of our head model which ensures robust novel view synthesis in Fig. 11. We also find noise occurred on the extracted surface, and this may be circumvented by using state-of-the-art NeRF based surface representation like [Wang et al. 2021; Yariv et al. 2021] and we leave it as a future work.

4.5 Bases Visualization

The semantical meaning of our hash table basis is inherited from the mesh blendshape used for tracking, see Sec. 3.3.1. Fig. 9 demonstrates some selected expression bases of our model and the mesh-based blendshape, where semantical correspondence between our bases and the mesh-based blendshape are shown. We can find that the expressions of our bases are consistent with the mesh-based

blendshape's expressions. Differently, the results generated by our model have higher rendering quality and highly personalized facial attributes. Moreover, our bases can represent various personalized facial details like hair and moles. The subject-specific habits, such as the wrinkles and folds movements, could also be seen on each basis of our model.

4.6 Novel View Synthesis

Our model also disentangles the camera parameters (Equ. (1)). Thus we can freely adjust the camera parameters of our model to generate target results with any desired rendering view. Fig 11 shows the novel view synthesis application based on our model. We first use a set of expression coefficients from the testset to combine bases to get the corresponding radiance field. Then the rendered images with different rendering views are generated by the volume rendering. Thanks to the 3D consistency of NeRF, these rendered results have remarkable multi-view consistency.

5 ABLATION STUDY

5.1 Discussion on Perceptual Loss

Fig. 12 shows the comparison between the results with/without perceptual loss supervision. It can be observed that the perceptual loss effectively improves the rendering quality and personalized facial attributes. This gain comes from the fact that the perceptual loss effectively maintains the visual similarity between the predicted image and the ground truth by minimizing the distance of the two images' features extracted by a pre-trained model.

5.2 Discussion on Expression-Aware Density Grid Update

The density grid could mark the spatial occupation of the density field and indicate the empty area to be skipped during raymarching, which could accelerate the computation and reduce hash collisions. Therefore, an effective grid update strategy is necessary for constructing our model. Unlike the original density grid from Instant NGP [Müller et al. 2022], where a static scene is considered, we need to handle a dynamic head scene. To this end, we consider more expressions of the training dataset and adopt an expression-aware

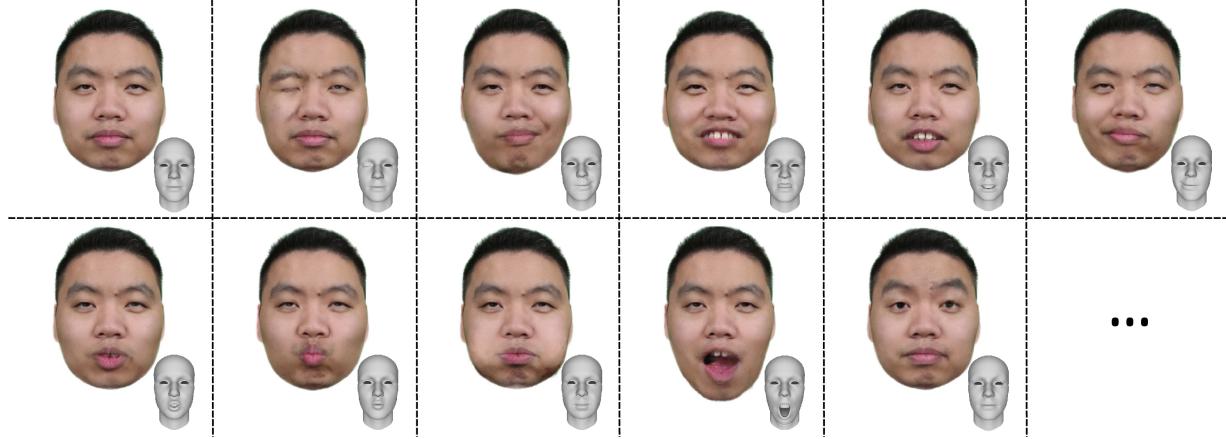


Fig. 9. Our bases are consistent with mesh blendshapes on the aspect of semantical meaning but with a photo-realistic rendering quality. Our bases contain much more subject-specific facial details.



Fig. 10. We extract the iso-surface from the density field with marching cubes. Our model could learn the geometric attributes like eyes, nose and detailed hair thanks to the radiance field background.

way to prune the density grid. As shown in Fig. 13, the proposed update strategy effectively helps our model generate more reasonable results. In contrast, it is difficult to deal with some expressions like mouth open if we only use a static neutral head to instruct density grid update. The training processes using/not using a density grid are shown in Fig. 14. We could see that a model could achieve better PSNR when a density grid is used.

6 LIMITATIONS

Similar with other NeRF based face modeling approaches like NerFACE and NHA, artifacts may occur in some local regions when extrapolating the expression coefficients to a value that is far from the training distribution. This problem might be circumvented by explicitly modeling the underlying geometry like some NeRF Editing approaches [Bao and Yang et al. 2022; Yuan et al. 2022], and we leave this as a future work.

The camera parameters and input conditions are important for NeRF based techniques. Large errors in tracking may cause losing details in our constructed model.

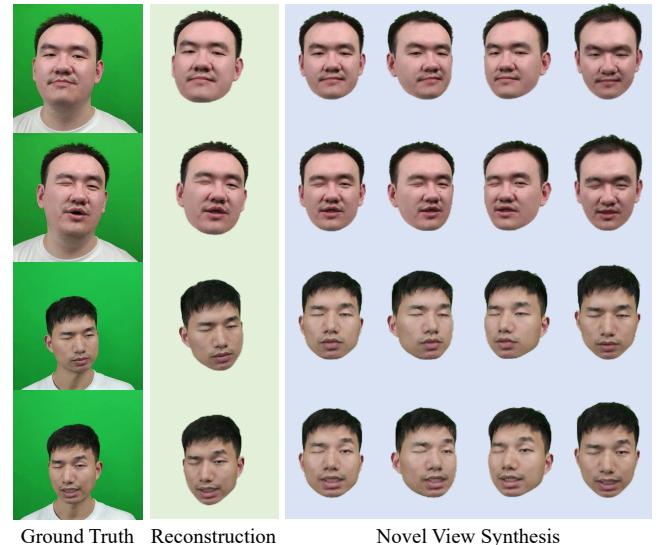


Fig. 11. Our model is a 3D consistent representation, and thus the view direction can be freely adjusted

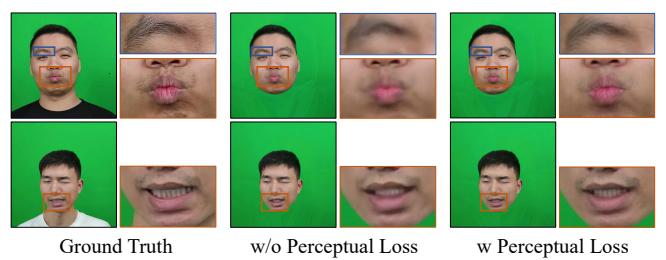


Fig. 12. Ablation study on Perceptual Loss

Thanks to the subject-specific NeRF training, our method works well for different genders, skin colors, accessories, and hairstyles.



Fig. 13. Ablation study on expression-aware density grid update strategy.

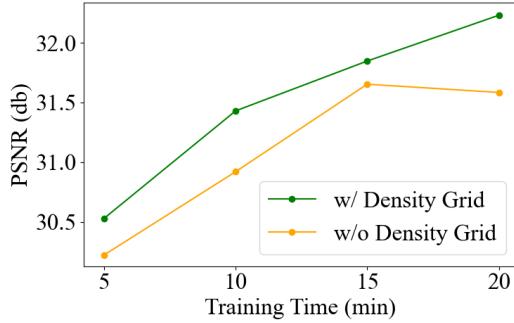


Fig. 14. Ablation study on training with/without density grid. Within the same training time, the rendering quality can be better with density grid.

However, if the hair is in a fast and heavy non-rigid deformation, artifacts may happen in hair region due to the lack of non-rigid deformation condition.

7 CONCLUSION

We have proposed a personalized semantic facial NeRF model, which could reconstruct a subject-specific 4D avatars with only minutes of monocular RGB video. Compared with mesh-based blendshape model, our model could generate photo-realistic results and model much more personalized facial attributes, such as hair, wearings, and even muscle movements. Meanwhile, our model has better representation ability and is easy for the MLP-based implicit function to learn the dynamic head. Compared with other NeRF-based parametric head models, our model can be constructed with much less time, has better rendering quality, and contains rich facial details. It is also worth pointing out that our implicit linear blending architecture can be adopted in other problems where a linear combination relationship plays an important role due to its representation ability and training efficiency. We believe our work is a forward step toward the future digital human.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (No.62122071, No.62272433), and the Fundamental Research Funds for the Central Universities (No. WK3470000021).

REFERENCES

- Ziqian Bai, Zhaopeng Cui, Xiaoming Liu, and Ping Tan. 2021. Riggable 3D Face Reconstruction via In-Network Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6216–6225.
- Bao and Yang, Zeng Junyi, Bao Hujun, Zhang Yinda, Cui Zhaopeng, and Zhang Guofeng. 2022. NeuMesh: Learning Disentangled Neural Mesh-based Implicit Field for Geometry and Texture Editing. In *European Conference on Computer Vision (ECCV)*.
- Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5855–5864.
- Volker Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. 187–194.
- Chen Cao, Menglei Chai, Oliver Woodford, and Linjie Luo. 2018. Stabilized real-time face tracking via a learned dynamic rigidity prior. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–11.
- Chen Cao, Qiming Hou, and Kun Zhou. 2014. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 1–10.
- Chen Cao, Yanlin Weng, Shun Zhou, Yiyi Tong, and Kun Zhou. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2013), 413–425.
- Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2016. Real-time facial animation with image-based dynamic avatars. *ACM Transactions on Graphics* 35, 4 (2016).
- Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16123–16133.
- Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*. 5799–5809.
- Bindita Chaudhuri, Noranart Vesdapunt, Linda Shapiro, and Baoyuan Wang. 2020. Personalized face modeling for improved face reconstruction and motion retargeting. In *European Conference on Computer Vision*. Springer, 142–160.
- Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. TensoRF: Tensorial Radiance Fields. In *European Conference on Computer Vision (ECCV)*.
- Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. 2021. Stereo radiance fields (srdf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7911–7920.
- Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. 2022. GRAM: Generative Radiance Manifolds for 3D-Aware Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8649–8658.
- Stephan J. Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien P. C. Valentini. 2021. FastNeRF: High-Fidelity Neural Rendering at 200FPS. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 14326–14335.
- Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2016. Reconstruction of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics (TOG)* 35, 3 (2016), 1–15.
- Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18653–18664.
- Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2022. StyleNeRF: A Style-based 3D Aware Generator for High-resolution Image Synthesis. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=iUzzTMUw9K>
- Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. 2018. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence* 41, 6 (2018), 1294–1307.
- Yudong Guo, Lin Cai, and Juyong Zhang. 2021a. 3D Face From X: Learning Face Shape From Diverse Sources. *IEEE Trans. Image Process.* 30 (2021), 3815–3827.
- Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. 2021b. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 5764–5774.
- Yang Hong, Bo Peng, Haiyan Xiao, Ligang Liu, and Juyong Zhang. 2022. HeadNeRF: A Real-time NeRF-based Parametric Head Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. 2017. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (ToG)* 36, 6 (2017), 1–14.
- Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3D avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)* 34, 4

- (2015), 1–14.
- Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. 2022. SelfRecon: Self Reconstruction Your Digital Avatar from Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5605–5615.
- H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. 2018. Deep Video Portraits. *ACM Transactions on Graphics (TOG)* (2018).
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- Hao Li, Thibaut Weise, and Mark Pauly. 2010. Example-based facial rigging. *ACM Transactions on Graphics (TOG)* 29, 4 (2010), 1–6.
- Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194–1.
- Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. 2022. Neural 3D Video Synthesis From Multi-View Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5521–5531.
- Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. 2020. Neural Sparse Voxel Fields. *NeurIPS* (2020).
- Xian Liu, Yinghao Xu, Qianyu Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. 2022. Semantic-Aware Implicit Neural Audio-Driven Video Portrait Generation. *arXiv preprint arXiv:2201.07786* (2022).
- Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–13.
- Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*. Springer, 405–421.
- Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* 41, 4 (July 2022), 102:1–102:15.
- Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. 2022. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5480–5490.
- Michael Niemeyer and Andreas Geiger. 2021. Giraffe: Representing scenes as compositional generative neural feature fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11453–11464.
- Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2021a. Nerfies: Deformable Neural Radiance Fields. *ICCV* (2021).
- Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. 2021b. HyperNerf: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields. *ACM Trans. Graph.* 40, 6, Article 238 (dec 2021).
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).
- Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2021a. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14314–14323.
- Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021b. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9054–9063.
- A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfelix, and F. Moreno-Noguer. 2019. GANimation: One-Shot Anatomically Consistent Facial Animation. (2019).
- Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. 2018. Generating 3D faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 704–720.
- Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinzhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance Fields without Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Katja Schwarz, Yili Liao, Michael Niemeyer, and Andreas Geiger. 2020. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems* 33 (2020), 20154–20166.
- Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.
- Robert W Sumner and Jovan Popović. 2004. Deformation transfer for triangle meshes. *ACM Transactions on Graphics (TOG)* 23, 3 (2004), 399–405.
- Cheng Sun, Min Sun, and Hwann-Tzong Chen. 2022. Direct Voxel Grid Optimization: Super-fast Convergence for Radiance Fields Reconstruction. (2022).
- Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020a. Neural voice puppetry: Audio-driven facial reenactment. In *European conference on computer vision*. Springer, 716–731.
- Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020b. Neural Voice Puppetry: Audio-driven Facial Reenactment. *ECCV* 2020 (2020).
- Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–12.
- J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. 2016. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*.
- Luan Tran and Xiaoming Liu. 2018. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7346–7355.
- Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. 2006. Face transfer with multilinear models. In *ACM SIGGRAPH 2006 Courses*. 24–es.
- Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *NeurIPS* (2021).
- Xueying Wang, Yudong Guo, Zhongqi Yang, and Juyong Zhang. 2022. Prior-Guided Multi-View 3D Head Reconstruction. *IEEE Trans. Multim.* 24 (2022), 4028–4040.
- Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
- Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. 2022. HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).
- Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 2020. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 601–610.
- Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. 2021. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems* 34 (2021).
- Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. 2021. i3dmm: Deep implicit 3d morphable model of human heads. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12803–12813.
- Alex Yu, Rui long Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. 2021a. PlenOctrees for Real-time Rendering of Neural Radiance Fields. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021b. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4578–4587.
- Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 325–341.
- Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. 2022. NeRF-Editing: Geometry Editing of Neural Radiance Fields. In *Computer Vision and Pattern Recognition (CVPR)*.
- Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. 2020. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*. Springer, 524–540.
- Juyong Zhang, Keyu Chen, and Jianmin Zheng. 2022. Facial Expression Retargeting From Human to Avatar Made Easy. *IEEE Trans. Vis. Comput. Graph.* 28, 2 (2022), 1274–1287.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018a. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018b. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.
- Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühlert, Xu Chen, Michael J Black, and Otmar Hilliges. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13545–13555.
- Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. 2021. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788* (2021).
- Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. 2021. MoFaNeRF: Morphable Facial Neural Radiance Field. *arXiv preprint arXiv:2112.02308* (2021).