Zero-shot Real Facial Attribute Separation and Transfer at Novel Views

Dingyun Zhang[1], Heyuan Li[2], and Juyong Zhang[1,*]

[1] University of Science and Technology of China, Hefei, Anhui, China
[2] The Chinese University of Hong Kong, Shenzhen, Guangdong, China
*corresponding author

Abstract. Real-time and zero-shot attribute separation of a given real-face image, allowing attribute transfer and rendering at novel views without the aid of multi-view information, has been demonstrated to be beneficial in real-world scenarios. In this work, we propose an alternating optimization framework and train it on attribute-blending (*i.e.*, unstructured) monocular images. Our framework leverages a pre-trained facial attribute encoder and a 3D-representation face synthesis decoder (*e.g.*, HeadNeRF) to reinforce and guide each other mutually. This allows the facial attribute encoder to better express and separate facial attributes and the face synthesis decoder to render faces with better image similarity and attribute consistency.

Keywords: neural rendering · alternating training · novel view synthesis · facial attribute transfer

1 Introduction

Real-time and zero-shot attribute separation of a given real face, along with attribute transfer and rendering at novel views without the aid of multi-view information, opens the door to a wide range of creative applications, such as talking face animation, face cloning and editing, training feature classifiers and generating synthetic images. In other words, it is desirable that a face avatar model could achieve a good balance in (1) **Zero-shot**, *i.e.*, for a test image, the model does not require optimization of network parameters or conditioned latent codes; (2) **Attribute transfer**, *i.e.*, for a test image, the model is capable of separating the attributes of the face into orthogonal spaces as much as possible and transferring a specific attribute, such as facial identity shape, expression, texture, illumination, hairstyle, and head pose, to another test face, without affecting the other facial attributes of the latter; (3) **Real-time**, *i.e.*, for a test image, the model completes the facial attribute separation and novel-view synthesis via an end-to-end forward pass; (4) **Realistic**, *i.e.*, the model can render facial appearance and expression details as rich as possible, rather than just rendering areas excluding hair, mouth interior and ears.

We investigate the research on neural face avatars and summarize previous works in Tab. 1. Explicit face models constructed from registered meshes have been widely used in modeling face avatars. However, due to the limitations of the overly simplistic principal component analysis method and the difficulty in obtaining real scans, most of these

Table 1: **A summary of current face avatar methods.** Δ_1 denotes that the facial attribute transfer could not guarantee a good attribute separation. Δ_2 denotes that the facial attribute transfer can only be performed on expression or head pose. Δ_3 denotes that the facial attribute transfer requires inputting a 3D scan, including the mesh. Δ_4 denotes that the model conducts novel-view synthesis of a real image via an end-to-end forward pass but is unable to separate the facial attributes.

Scheme	Methods	Zero-shot	Transfer	Real-time	Realistic
Explicit 3D Models	[4,43,24,74,86]		✓		
	[18,21,12]	✓	✓	✓	
3D-aware GANs	[10,17,11,28,52]				✓
	[15]	✓		Δ_4	✓
	[71,69,16,63,77]		Δ_1		✓
Personalized Avatars	[22,54,87,2,27]		Δ_2		✓
Talking Head	[60,19,45,35,64]		Δ_2		✓
Implicit Face Models	[32,89]		✓		✓
	[23]		✓		
	[81,76]		Δ_3		✓
	Ours	✓	✓	✓	✓

methods can only model and render the facial region, excluding the hair, mouth interior, and ears. Recent 3D-aware GANs [26] using implicit representations or StyleGAN-based methods can synthesize realistic faces. However, most of these methods require time-consuming GAN inversion for a real face image. For their synthesized fake face images, some models [10,17,11,28,52] are unable to separate and transfer facial attributes, while most of others [71,69,16,63,77] exhibit visible incompleteness in separating certain attributes (Fig. 1). Personalized avatars and talking head methods are often trained in a person-specific manner and can only separate facial expressions and head pose attributes. Recent models using implicit representations either require optimizing latent attribute codes during testing [32,89,23] or rely on mesh input for model fitting [81,76], thereby limiting their generalization ability to unseen identities and expressions.

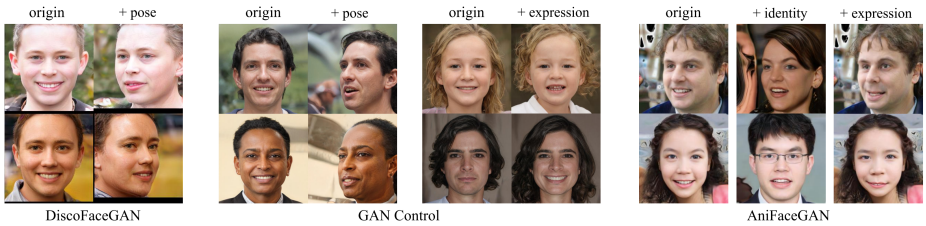

Fig. 1: **Recent controllable 3D-aware GAN methods have shown limitations over their synthesized fake faces in separating a certain facial attribute from others.** 'origin' refers to the face image synthesized from a random code by the corresponding model. '+' means changing a specific attribute code from the original random code. DiscoFaceGAN [16], for instance, does not achieve thorough separation, as controlling for head pose noticeably impacts the expression. GAN Control [63], when manipulating head pose or expression over the original image, significantly affects other facial attributes. AniFaceGAN [77] exhibits a significant influence of identity (*i.e.*, facial identity shape and texture) attribute on the expression attribute, whereas controlling for expression sometimes affects identity.

In this paper, inspired by the face reconstruction works [20,42], the learning of the facial attribute encoder to separately parameterize the attributes from a real image and the adaptation of the 3D-representation face synthesis decoder to render better a face image based on the conditioned codes is solved jointly using an expectation-maximization-like [13] procedure, where we train the two networks in an alternating manner. The motivation for doing so is based on the observation that during the adaptation of the face synthesis decoder, it optimizes the attribute labels of the face, which could guide the facial attribute encoder to promote the semantic expressiveness of attribute parameter prediction. Conversely, attribute parameter representation of a face with better expressiveness and separation can, in turn, serve as better initialized conditioned codes for the 3D-aware decoder to render a face image that has image similarity and attribute consistency with the ground truth image. Thus, both aspects can be considered as mutually dependent, similar to a chicken-and-egg relationship.

In our task, we construct the facial attribute encoder based on a face recognition network, face reconstruction networks, and a hairstyle encoding network and pre-train them. HeadNeRF [32], a 3D-aware face model based on neural implicit representation, is chosen as the face synthesis decoder. We use these models as an example of our alternating training approach in enhancing the facial attribute representation and separation capability of the encoder and the rendering quality of the face synthesis decoder. The alternating training in each round consists of two steps. In the first step, we update the network parameters and conditioned attribute labels of the face synthesis decoder, while in the second step, we update the parameters of the facial attribute encoder. We only train our model on attribute-blending (*i.e.*, unstructured) and non-multi-view 2D in-the-wild datasets. Considering that lacking the aid of multi-view information for a single identity can significantly degrade the high-frequency rendering quality of neural radiance field (NeRF) [34], we incorporate a pre-trained blind face restoration network, DifFace [84], as a refinement network during the inference stage. This addition aims to enhance the rendering quality of the face synthesis decoder, making the rendering results for the face images more realistic. In order to ensure fairness, we do not utilize refined images during experimental testing. Instead, we qualitatively showcase them as references. Inspired by [16,19,71,63], we employ both self-supervised disentanglement loss and cycle-consistency loss as part of the alternating training. Through relevant experiments, we demonstrate the results of the proposed method. In summary, our contributions can be summarized as follows:

- We extend the alternating training algorithm to the focus that to enhance the ability of the facial attribute encoder in representing and separating attributes, and to improve the rendering quality of the 3D-aware face model.
- We present a model that could realize real-time and zero-shot attribute separation of a given real face, allowing attribute transfer and rendering at novel views without the aid of multi-view information.
- We demonstrate the proposed method through relevant experiments.

2 Related Works

2.1 Explicit Face Morphable Models

Explicit representation is widely used for 3D face modeling. It is typically built by performing Principal Component Analysis (PCA) on numerous registered 3D facial scans and represents a 3D face as the linear combination of a set of orthogonal bases. Blanz and Vetter [4] first introduced the concept of a 3D Morphable Face Model (3DMM). Since then, many efforts [1,5,6,9,24,86] have been devoted to improving the performance of 3DMM by either improving the quality of captured face scans or the structure of 3D face model. However, acquiring registered 3D data is laborious, and most of the existing methods [4,5,6,9,24,43,58] can only render the texture of the facial region, excluding the hair, mouth interior, and ears. Meanwhile, the rendered faces produced by these methods often exhibit visible differences in identity or expression compared to the original faces, resulting in the sense of artificiality. In addition, most models are optimization-based for fitting a real image, requiring solving the inverse rendering equation and, therefore, not real-time.

Recent state-of-the-art regression-based methods [18,21,12] typically render face images with estimated illumination, texture, and geometry of the face model using a differentiable renderer [59,44] and compare the synthetic images with the inputs. Such an analysis-by-synthesis strategy facilitates the demand for in-the-wild images and help to recover geometric details. However, their separation of attributes is visibly incomplete, as changing the parameter of one attribute would significantly affect other facial attributes of the rendered face.

2.2 3D-aware Implicit Models

3D-aware methods aim to learn a model that can explicitly control the camera viewpoints of the synthesized content. Neural implicit functions have been used in numerous works [53,46,68,67,47,51] to represent 3D scenes or faces. In contrast to explicit representations (*e.g.*, meshes or voxel grids), neural implicit representation is well-suited to model complex surfaces and realistic textures. Recent advances in 3D-aware GANs [26] have enabled the synthesis of realistic multi-view fake faces. Some of these approaches [11,10,17,52,28,50] utilize neural implicit representations but do not focus on separating facial attributes. Additionally, rendering novel views of real images requires time-consuming GAN inversion [61,82] to optimize the input codes. Very recently, some work [15] trained an encoder for the GAN [17] to map a real image to the corresponding latent code. However, it does not address the limitation of real-time separation of facial attributes and attribute transfer. Some implicit [77,69,41,62,72] or 2D-based [16,63,71,49,8,38,25,55] 3D-aware controllable GANs incorporate 3DMM priors to achieve attribute separation control of generated fake faces. However, as shown in Fig. 1, these models often exhibit visible deficiencies in attribute separation control.

Recent works focused on rendering animatable personalized avatars [22,54,87,2,90,27,3,88] or talking head animation [30,45,19,60,35,64,65,79] often need to train a specific model for one or two persons from monocular videos and can only separate facial expressions and head pose attributes. Other works [75,57] could

render static personalized avatars from multi-view images with high fidelity but could not separate facial attributes.

[32] propose the first 3D-aware NeRF-based [34] parametric face model, which controls the facial identity shape, expression, texture, illumination, and head pose of the rendered face by corresponding latent codes. [89,23] propose a model in a similar way. [89] is unable to render the hair region and control the illumination. [23] is incapable of rendering the hair, mouth interior, and ear regions, and it does not further divide the identity attribute into facial identity shape and texture attributes. Although these models enable identity and expression editing by adjusting the associated 3DMM parameters, the limited representation ability of latent parameters bound their ability to recover some facial details in the original frames and their generalization ability to unseen identities, expressions, and head poses. Moreover, to fit a real face image, these methods require time-consuming optimization for the initialized latent attribute codes. [81] propose i3DMM, a deep implicit 3D morphable model that can be animated by learned latent codes. [76] define the deformation filed by standard linear blend skinning (LBS), which allows the avatars to be directly animated by FLAME parameters. However, to fit a real face image, both methods require the simultaneous acquisition of the face image and its corresponding mesh to perform latent attribute code optimization and render the face.

2.3 Disentanglement Representation Learning

Disentangled representation learning (DRL) for face images has been vividly studied in the past. Compared to the real-time attribute separation of a real face, most 3D-aware controllable GANs emphasize seeking an interpretable and highly disentangled latent space of the generator, allowing for explicit control over the facial attributes of the synthesized fake faces. A common tactic is to hallucinate or render synthetic images varying in different attributes and then jointly learn the attribute differences from these images. [49] disentangles head pose and identity with unsupervised learning using 3D convolutions and rigid feature transformations. [16] proposes imitative-contrastive learning to mimic the 3DMM rendering process by the generative model. A similar strategy has also been adopted with concurrent and follow-up works [25,55,71,70,8]. [38] uses a custom 3D image rendering pipeline to generate an annotated synthetic dataset. This dataset is later used to acquire controls matching the synthetic ground truth, allowing [38] to add controls. [63] utilizes a pairwise contrastive loss to understand the positive and negative relationships between synthetic training pairs for different attribute spaces. One-shot talking head model [19] employs a similar contrastive learning strategy to separate expression and head pose from other facial attributes.

Following [16], the implicit representation [77,69] mimics mesh deformation to achieve direct control of the identity, expression, and head pose. [72] explicitly models the deformation fields to enforce the disentanglement between geometry (*i.e.*, identity shape and expression) and appearance (*i.e.*, texture and illumination). [32,89,80,76,23] rely on attribute-disentangled multi-view annotated datasets to learn the attribute separation of the latent space, with the training data for the first five models being collected professionally in a laboratory setting. For a real face image, some regression-based face reconstruction methods [18,21] predict the parameters of facial attributes through

end-to-end unsupervised training or further incorporating the designed consistency losses.

3 Method

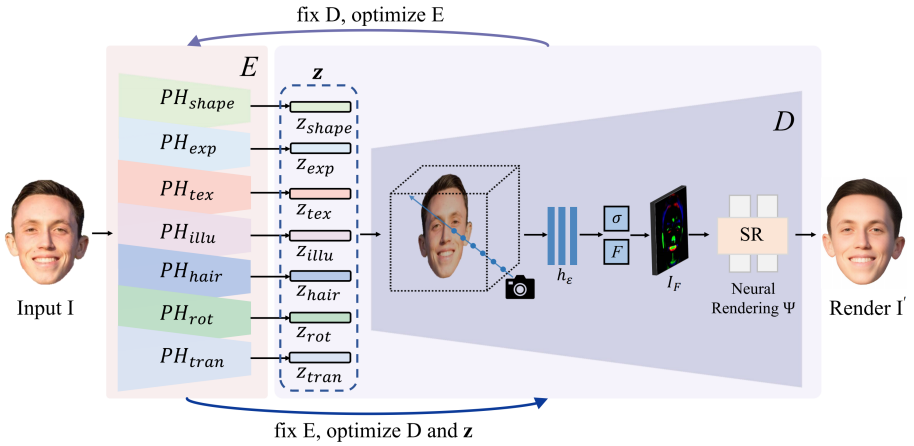

Fig. 2: **Method overview.** Our model consists of a facial attribute encoder, E, composed of seven facial attribute prediction heads (PH), and a face synthesis decoder, D. E takes in a given face image I and projects it into the latent space divided into separate attribute sub-spaces, generating 1-D feature vector $\mathbf{z}$. The conditioned attribute codes $\mathbf{z}$ are fed to the volumetric-representation face synthesis decoder D to render a reconstruction I'. We alternately trained the decoder D and encoder E using an EM-like heuristic algorithm, enabling them to synergize and provide each other with informative guides or priors. See text in Sec. 3 for details.

First, in Sec. 3.1, we present the model architecture that can perform zero-shot facial attribute separation and transfer at novel views from a real face image after training. Then, in Sec. 3.2, we introduce how to train our model using an EM-like heuristic training algorithm. In Sec. 3.3, we show the model parameters initialization of the EM-like alternating training procedure. Finally, in Sec. 3.4, we demonstrate how to refine the rendered face image using an additional pre-trained face restoration network and clarify its role in our paper.

3.1 Model Architecture

As illustrated in Fig. 2, our model consists of a facial attribute encoder E that takes in a face image I and projects it into the latent space divided into separate attribute sub-spaces, generating a 1-D feature vector $\mathbf{z} = [\mathbf{z}_{\text{shape}}, \mathbf{z}_{\text{exp}}, \mathbf{z}_{\text{tex}}, \mathbf{z}_{\text{illu}}, \mathbf{z}_{\text{hair}}, \mathbf{z}_{\text{rot}}, \mathbf{z}_{\text{tran}}]$ about attribute facial identity shape, expression, texture, illumination (lighting) under the Spherical Harmonics illumination model [56], hairstyle (hair shape), face pose rotation and translation under the standard perspective camera model for projecting a point in 3D space onto the image plane, respectively. The feature vector $\mathbf{z}$ is fed as a condition code to a volumetric-representation face synthesis decoder D to render a reconstruction I'.

Facial Attribute Encoder. As illustrated in Fig. 2, the facial attribute encoder E comprises seven prediction heads: PH_{shape} for identity shape, PH_{exp} for expression, PH_{tex} for texture, PH_{illu} for illumination, PH_{hair} for hairstyle, PH_{rot} for face pose rotation, and PH_{tran} for face pose translation. The PH_{shape} utilizes Adaface [36], a face recognition network with ResNet50 [31] as its backbone. The prediction heads PH_{exp}, PH_{tex}, PH_{illu}, PH_{rot}, and PH_{tran} all employ single image reconstruction network, R-Net, from Deep3DFace [18]. Many previous methods [32,89,23,18,21,12] for parameterizing and rendering a real face image at novel views did not consider a specific representation and rendering of the hairstyle. Recently, there have been methods [76,81] that collect or utilize full photogrammetric attribute-disentangled head scans, including mesh, for training purposes, enabling the inclusion of the hair component during rendering. However, GANHead [76] is unable to parameterize hairstyle through latent attribute code, and the hairstyle latent code of i3DMM [81] can only take discrete values - short, long, cap1, or cap2, which to some extent restricts the expressiveness of hairstyle latent code in rendering the real face hairstyle. We adopt the shape encoding network of the 2D hair editing GAN CtrlHair [29] as the predicting head PH_{hair} for hairstyle, which allows us to avoid using professional handcrafted 3D face scan data and instead train on a large number of easily accessible unstructured 2D face images. Although CtrlHair's shape encoding network cannot attribute separate the hairstyle and head pose, meaning that the feature vector obtained from the network for the same hairstyle under different head poses often has significant differences, our experiments demonstrate that after training the entire model, this entanglement can be canceled out. The dimensions of the latent attribute codes are as follows: $\mathbf{z}_{\text{shape}} \in \mathbb{R}^{512}$, $\mathbf{z}_{\text{exp}} \in \mathbb{R}^{64}$, $\mathbf{z}_{\text{tex}} \in \mathbb{R}^{80}$, $\mathbf{z}_{\text{illu}} \in \mathbb{R}^{27}$, $\mathbf{z}_{\text{hair}} \in \mathbb{R}^{16}$, $\mathbf{z}_{\text{rot}} \in \mathbb{R}^{3}$, and $\mathbf{z}_{\text{tran}} \in \mathbb{R}^{3}$, where rotation is defined using Euler angles.

Face Synthesis Decoder. The face synthesis decoder D utilized is HeadNeRF [32], a model that integrates 3DMM with the NeRF representation and is capable of synthesizing 3D-aware faces conditioned on 3DMM attributes - identity shape, expression, texture, illumination, and head pose. The modification we made was to adjust the dimensions of the conditioned attribute code to match the output dimensions of the facial attribute encoder E instead of using the previous dimension of HeadNeRF, which was set to facilitate initializing the latent codes with the solution obtained by solving inverse rendering optimization based on [74]. We additionally include the hairstyle attribute code $\mathbf{z}_{\text{hair}}$ into the conditioned latent codes of HeadNeRF.

Next, we briefly introduce the architecture of the face synthesis decoder D that we have employed. D is a NeRF-based parametric model, which can render an image I' with specified attributes for the given condition codes. It is formulated as: $\text{I}' = \text{D}(\mathbf{z}_{\text{shape}}, \mathbf{z}_{\text{exp}}, \mathbf{z}_{\text{tex}}, \mathbf{z}_{\text{illu}}, \mathbf{z}_{\text{hair}}, \mathbf{z}_{\text{rot}}, \mathbf{z}_{\text{tran}})$, where $\mathbf{z}_{\text{rot}}$ is then transformed to a rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times3}$. The MLP-based implicit neural function h_ϵ of NeRF is formulated as:

$$h_\epsilon : (\gamma(\mathbf{x}), \mathbf{z}_{\text{shape}}, \mathbf{z}_{\text{exp}}, \mathbf{z}_{\text{tex}}, \mathbf{z}_{\text{illu}}, \mathbf{z}_{\text{hair}}) \mapsto (\sigma, F), \tag{1}$$

where ϵ represents the network parameters, $\gamma(\cdot)$ is the positional encoding in NeRF [34], and $\mathbf{x} \in \mathbb{R}^3$ is a 3D point sampled from one casted camera ray. h_ϵ outputs the density value σ at $\mathbf{x}$ and an intermediate feature vector $F(\mathbf{x}) \in \mathbb{R}^{256}$. Then the 2D feature map

$I_F \in \mathbb{R}^{256\times32\times32}$ is obtained by performing the volume rendering:

$$I_F(r) = \int_0^\infty w(t) \cdot F(r(t))dt, \tag{2}$$

where $w(t) = exp(-\int_0^t \sigma(r(s))ds) \cdot \sigma(r(t))$ and $r(t)$ is a ray emitted from the camera center. I_F then passes through a 2D neural rendering network Ψ whose design concept is inspired by StyleNeRF [28], progressively increasing its resolution, and eventually be transformed into the rendered image $\mathrm{I}' \in \mathbb{R}^{3\times1024\times1024}$.

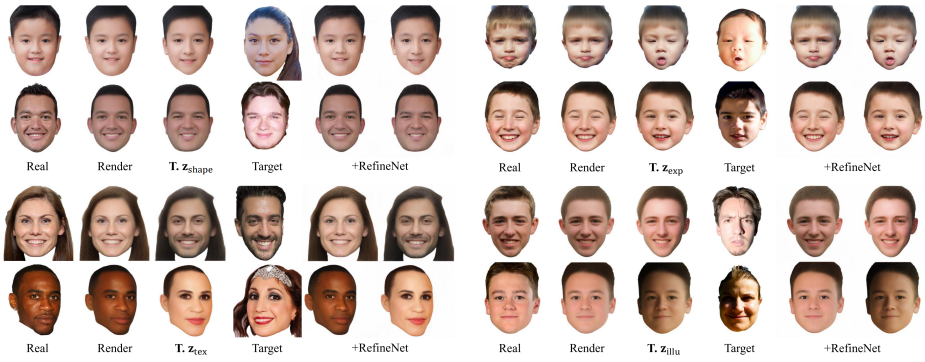

Fig. 3: **Zero-shot attribute separation from a real image**. 'Real' and 'Target' respectively represent the source and target ground truth real-face images. 'Render' represents the rendering result on the source image. '**T.** $\mathbf{z}_*$' denotes replacing the corresponding attribute code of the source image with that of the target image. '+RefineNet' means the results for the rendered images refined by the adopted face restoration network. Our model is capable of real-time attribute-separated representation and rendering a real-face image. The facial attribute encoder E separates various attributes of the face into an orthogonal latent space as much as possible and accurately represents them using latent codes. This allows the face synthesis decoder D to render a face that resembles the attributes of the real face image. During attribute transfer, the rendered face maintains similarity to the target face in the transferred attributes, while the unmodified attributes are preserved well. In this figure, we present examples of transferring identity shape, expression, texture, and illumination. See text in Sec. 4.2 for details.

3.2 EM-like Alternating Training Procedure

Due to the mutual dependencies between the facial attribute encoder E and face synthesis decoder D, we employ an EM-like heuristic training strategy, where we train the two networks in an alternating manner. Similar to other EM-like training strategies, our training process starts from a rough initialization of the model parameters (as described in Sec. 3.3). We then alternately optimize the face synthesis decoder D and facial attribute encoder E, as described in the following.

Training the face synthesis decoder. When training the face synthesis decoder D, the parameters of the facial attribute encoder E are fixed, and only D and the conditioned attribute codes $\mathbf{z}$ are optimized. At this step, we assume that the facial attribute encoder E is already good enough, meaning it can separate and parameterize the facial attribute

of a real image into the latent space orthogonally as much as possible, and the resulting latent attribute code $\mathbf{z}$ exhibits sufficient expressiveness for facial characteristics.

The overall objective function of this step is:

$$\mathcal{L}_{\text{D}} = \mathcal{L}_{\text{pix}} + \mathcal{L}_{\text{perc}} + \mathcal{L}_{\text{id}} + \mathcal{L}_{\text{reg}}. \tag{3}$$

The photometric consistency term $\mathcal{L}_{\text{pix}}$ is a pixel-wise L1 distance measured between the synthesized image I' and the ground truth image I, which is formulated as:

$$\mathcal{L}_{\text{pix}} = \frac{1}{|\text{M} \odot \text{I}|} \|\text{M} \odot (\text{I}' - \text{I})\|_1. \tag{4}$$

M is the head region mask of I and $\odot$ stands for a pixel-wise Hadamard product operator.

The perception-level loss $\mathcal{L}_{\text{perc}}$ measures perceptual and semantic differences between two images with an image classification network ϕ:

$$\mathcal{L}_{\text{perc}} = \sum_{i=1}^{5} \frac{1}{|\phi_i(\text{I})|} \|\phi_i(\text{I}) - \phi_i(\text{I}')\|_1, \tag{5}$$

where i denotes the i-th layer of VGG19 [66] network pre-trained on ImageNet [39].

The face identity loss $\mathcal{L}_{\text{id}}$ is the cosine distance between the embeddings of a pre-trained face recognition network f [14]:

$$\mathcal{L}_{\text{id}} = 1 - \frac{f(\text{I}) \cdot f(\text{I}')}{\|f(\text{I})\|_2 \|f(\text{I}')\|_2}. \tag{6}$$

We use this loss to ensure that the rendered image I' looks like the same person as the ground truth subject.

Finally, we adopt a latent space regularization loss $\mathcal{L}_{\text{reg}}$ to prevent facial attribute degeneration:

$$\mathcal{L}_{\text{reg}} = \sum_{*} \omega_* (1 - \frac{\mathbf{z}_* \cdot \mathbf{z}_*^0}{\|\mathbf{z}_*\|_2 \|\mathbf{z}_*^0\|_2}), \tag{7}$$

where $\mathbf{z}_*^0$ denotes the initial values of the seven attribute codes obtained from the facial attribute encoder E, and ω_* represents the loss weight.

During training, the face synthesis decoder D updates the seven conditioned attribute codes $\mathbf{z}$, which can be regarded as the updated labels obtained through D. This step of training aims to encourage the attribute codes to fall in more semantically meaningful and attribute expressive location in the latent space while improving the consistency between the rendered face I' synthesized by face synthesis decoder and real face I, thereby enhancing the decoder's ability to represent a real face based on conditioned attribute codes.

Training the facial attribute encoder. In the second step, we continue to optimize the parameters of the facial attribute encoder E while keeping the face synthesis decoder D fixed. At this step, we assume that the face synthesis decoder D and the updated latent attribute codes $\mathbf{z}^{\text{D}}$ obtained from the previous step are already good enough, meaning

that D's latent space is smooth and meaningful, and the updated attribute codes have enough expressiveness for the ground truth face image, thereby could serve as labels for guiding the optimization of the facial attribute encoder E. The overall objective function of this step is:

$$\mathcal{L}_{\mathrm{E}} = \lambda_{\mathrm{cod}}\mathcal{L}_{\mathrm{cod}} + \lambda_{\mathrm{cyc}}\mathcal{L}_{\mathrm{cyc}} + \lambda_{\mathrm{dis}}\mathcal{L}_{\mathrm{dis}} \tag{8}$$

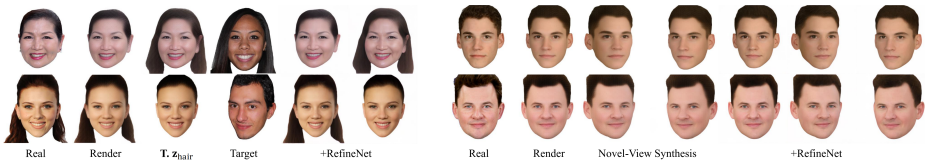

Fig. 4: **Zero-shot attribute separation from a real image and novel-view synthesis**. On the left side of this figure, we present examples of transferring the hairstyle. On the right side of this figure, we illustrate examples of synthesizing novel views of the rendered image from a single image by changing the pose. See text in Sec. 4.2 for details.

Code Consistency Loss. The code consistency loss $\mathcal{L}_{\mathrm{cod}}$ is defined as the separate cosine distance between the parameterized attribute codes obtained by the encoder E for the ground truth image I and the seven updated conditioned attribute codes $\mathbf{z}_*^{\mathrm{D}}$ obtained after the previous training step:

$$\mathcal{L}_{\mathrm{cod}} = \sum_{*} \overline{\omega}_* \left(1 - \frac{\mathrm{PH}_*(\mathrm{I}) \cdot \mathbf{z}_*^{\mathrm{D}}}{\|\mathrm{PH}_*(\mathrm{I})\|_2 \|\mathbf{z}_*^{\mathrm{D}}\|_2}\right). \tag{9}$$

$\overline{\omega}_*$ represents the loss weight. We use the updated condition codes as labels to encourage the facial attribute encoder E in separating the attributes of real image I to more meaningful and expressive locations in the latent space.

Cycle-consistency Loss. The cycle-consistency loss computes the separate difference between the attribute codes of the ground truth image I predicted by the encoder E and those of the rendered image I':

$$\mathcal{L}_{\mathrm{cyc}} = \sum_{*} \overline{\omega}_* \left(1 - \frac{\mathrm{PH}_*(\mathrm{I}) \cdot \mathrm{PH}_*(\mathrm{I}')}{\|\mathrm{PH}_*(\mathrm{I})\|_2 \|\mathrm{PH}_*(\mathrm{I}')\|_2}\right). \tag{10}$$

We use this loss to encourage the encoder E to predict more stable latent codes for facial attributes, and the rendered face I' conditioned on these latent codes should convey the corresponding attribute content of the input image I.

Disentanglement Loss. Following the concept in [16,71,70], given two latent attribute codes $\mathbf{z}_i = [\mathbf{z}_{a1}^i, ..., \mathbf{z}_{an}^i]$, $\mathbf{z}_j = [\mathbf{z}_{a1}^j, ..., \mathbf{z}_{an}^j]$ predicted by the attribute encoder E from the corresponding image I^i, I^j in a training batch, we randomly vary one attribute code of $\mathbf{z}_j$ while keeping others unchanged. By replacing $\mathbf{z}_{ak}^j$ with $\mathbf{z}_{ak}^i$, we obtain a new attribute code $\hat{\mathbf{z}_j}$. $\mathbf{z}_j$ and $\hat{\mathbf{z}_j}$ differ only at sub-code for attribute ak, and share the same

sub-codes for attribute $al, \forall l \neq k$. For example, ak can represent facial identity shape. In an ideal situation, $\hat{\mathbf{z}_j}$ should retain the expression, texture, scene illumination, hairstyle, and head pose of $\mathbf{z}_j$, but should perform the identity shape specified in $\mathbf{z}_i$. $\hat{\mathrm{I}^j} = \mathrm{D}(\hat{\mathbf{z}_j})$ corresponding to the image I^j should be modified according to the sub-code of $\mathbf{z}_i$.

Since we do not have ground truth for such a variation, *i.e.*, the image $\hat{\mathrm{I}^j}$ is unknown, we employ supervision based on the disentanglement loss $\mathcal{L}_{\text{dis}}$. The decoded image $\hat{\mathrm{I}^j}$ is again passed through the attribute encoder E to generate $E(\hat{\mathrm{I}^j})$. The disentanglement loss $\mathcal{L}_{\text{dis}}$ enforces that $E(\hat{\mathrm{I}^j})$ should have the same identity shape code as $\mathbf{z}_i$ and enforces consistency of the parameters that should not be changed by the performed edit operation. In the case of modifying identity shape values, the parameters that should not change are expression, texture, illumination, hairstyle, and head pose parameters. This leads to the disentanglement loss function:

$$\begin{aligned}\mathcal{L}_{\text{dis}} = \overline{\omega}_{ak} \left(1 - \frac{\mathbf{z}^i_{ak} \cdot \mathrm{PH}_{ak}(\hat{\mathrm{I}^j})}{\|\mathbf{z}^i_{ak}\|_2 \|\mathrm{PH}_{ak}(\hat{\mathrm{I}^j})\|_2}\right) \\ + \sum_{l \neq k} \overline{\omega}_{al} \left(1 - \frac{\mathbf{z}^j_{al} \cdot \mathrm{PH}_{al}(\hat{\mathrm{I}^j})}{\|\mathbf{z}^j_{al}\|_2 \|\mathrm{PH}_{al}(\hat{\mathrm{I}^j})\|_2}\right).\end{aligned} \tag{11}$$

We perform the same operations in reverse order, *i.e.*, in addition to replacing $\mathbf{z}^j_{ak}$ of $\mathbf{z}_j$ with $\mathbf{z}^i_{ak}$, we also replace $\mathbf{z}^i_{ak}$ of $\mathbf{z}_i$ with $\mathbf{z}^j_{ak}$ and obtain a new attribute code $\hat{\mathbf{z}_i}$. The corresponding disentanglement loss will be calculated in the same way.

3.3 Model Parameters Initialization

As with every other EM-like training strategy, our training needs a proper initialization of the model parameters. To provide an initialization for the prediction heads of the facial attribute encoder E, we individually pre-trained the adopted prediction heads based on their official implementations, ensuring that they can provide certain semantic priors and attribute separability. After initializing the facial attribute encoder E, we trained the face synthesis decoder D for 10 epochs on the training set of the FFHQ dataset [33] using the method described in Sec. 3.2. This step was taken to allow decoder D to initially learn to understand the semantic information conveyed by conditioned attribute codes and acquire the ability to synthesize face images. Initially, the optimization of decoder D and conditioned latent codes will bring significant changes to the latter. Therefore, we then train the facial attribute encoder E using only the code consistency loss $\mathcal{L}_{\text{cod}}$ for 5 epochs to synchronize the semantics between E and D, ensuring the stability of subsequent formal training. Initializing the decoder's parameters only on the FFHQ training set may lead to instability during formal training on the complete mixed training set due to domain differences between datasets. However, since we pre-aligned the training data, such instability only appeared in a small number of images, and we excluded these images.

3.4 Rendering refinement with blind face restoration

In our paper, we use HeadNeRF [32] as an instance of the face synthesis decoder D, which mitigates the high computational cost of NeRF by first rendering low-resolution feature maps and then applying 2D CNNs for super-resolution. However, this structure suffers from a common issue of losing image details, possibly due to the black-box rendering of CNNs. Another major issue is that the NeRF architecture is suitable for novel-view rendering from multi-view images, but we use single-view attribute-blending training data, which is easier to obtain. Therefore, although HeadNeRF renders rich details on the multi-view data used in its work, the rendering effect on single-view images lacks high-frequency details, such as the texture of hair and fur, due to the lack of auxiliary multi-view information, resulting in overly smoothed rendered heads. On the other hand, it's common to apply refinement networks on top of the rendered images to generate more realistic texture details [89,78]. From the visual performance standpoint, we alleviate this limitation by employing a pre-trained real-time blind face restoration network called DifFace [84]. We feed the face I' rendered by our model into DifFace, which outputs refined image I'-refine of the same resolution. In order to ensure fairness, we do not utilize the refined images during experimental testing. Instead, we qualitatively showcase them as references.

4 Experiment

4.1 Implementation Details

Datasets and Data Pre-processing. **FFHQ** [33] and **CelebAMask-HQ** [40] datasets contain 70,000 and 30,000 in-the-wild single-view face images respectively, with rich identity and age diversity, and high image resolution (1024×1024 for the former and 512×512 resolution for the latter). **AffectNet** [48] is a large-scale emotion dataset with face images acquired from the Internet, covering seven emotional states (*i.e.*, anger, disgust, fear, happiness, neutral, sadness, and surprise). To maintain consistency, we resize all the images to 1024×1024. To stabilize the training, we align each image to a similar center before training. We use an off-the-shelf semantic segmentation network [83] to obtain the segmentation labels of each image and remove images that contain hats, earrings, and necklaces. Additionally, we employed these segmentation labels to replace the background region (*i.e.*, without a head) of the images with a white backdrop and generate the head region masks for the images. We randomly take approximately 320,000 images from these datasets as the training set and evaluate the model on the randomly selected images that were not included in the training set.

Training Details. We use PyTorch to implement our model. For the face synthesis decoder D, 1024 rays are sampled in an iteration, each with 64 sampled points in the coarse volume. Similar to [32], we remove the hierarchical volume sampling of NeRF to speed up training and inference. In the formal EM-like training, We first train decoder D for 5 epochs, then train encoder E for 5 epochs, and repeat this process alternately. In order to ensure consistency with the optimizers used during the model parameters

initialization, seven Adam optimizers [37] are used for training the face synthesis decoder D and the prediction heads excluding PH_{shape} respectively, while a SGD optimizer [7] is used to optimize PH_{shape}. The initial learning rate of the Adam optimizer was set to 5×10^{-4}, and that of the SGD optimizer was set to 0.1. When training the facial attribute encoder, the gradient of each attribute component of the objective terms in $\mathcal{L}_{\text{E}}$ is summed and backpropagated to the corresponding prediction head, and the weights of the prediction heads are adjusted based on the propagated gradient, respectively. The whole training is conducted on 5 NVIDIA RTX3090 GPUs for 150 epochs.

Fig. 5: **Visual comparison of representation ability.** 'Ours-refine' means the results of our rendered images refined by the adopted face restoration network. 'MoFaNeRF-fine' means its refined results. For a real image, we show the regressed prediction [18,21] or fitted predition [32,73] of the baseline models. The better attribute consistency between the rendered face image and the ground truth real image could indicate a better representation ability of the latent attribute codes. See text in Sec. 4.3 for details.

4.2 Zero-shot Attribute Separation From Single Image

We validate whether our facial attribute encoder E can zero-shot separate facial attributes from a real face image as orthogonally as possible to the latent space of the face synthesis decoder D. We verify this through facial attribute transfer experiments.

As shown in Fig. 3, we use encoder E to predict facial attribute codes $\mathbf{z}$ for both the source ('Real') and target ('Target') real face images. We replace one specific attribute code $\mathbf{z}_*$ of the source image with the counterpart from the target image. The original and modified conditioned attribute codes are fed into the face synthesis decoder D for rendering, denoted as 'Render' and '**T.** $\mathbf{z}_*$' respectively for the rendered results, where '**T.**' means 'Transfer'. The desired outcome is that the rendered face 'Render' should also strive to be as similar as possible to the source real image 'Real' across various facial attributes. Simultaneously, '**T.** $\mathbf{z}_*$' should exhibit sufficient consistency to the target face in terms of the modified attribute $\mathbf{z}_*$ while the remaining unmodified attributes should be well preserved. Similarity reflects the expressive capacity of attribute encoder E in predicting latent codes for real facial attributes, while invariance demonstrates the good separation between the latent codes corresponding to different attributes. The results in Fig. 3 thus reveal that our model demonstrates good expressive and attribute separation capacity for facial identity shape, expression, texture, and illumination from a real image without the assistance of multi-view information in both training and testing.

As mentioned in Sec. 3.1, it is difficult to continuously parameterize the hairstyle attribute, especially without the aid of a professional multi-view 3D scan training dataset that includes hair. The examples on the left side of Fig. 4 demonstrate that the facial attribute encoder E's prediction head for hairstyle PH_{hair} can express and separate this attribute effectively. The hairstyle transfer results from a different head pose of the target image also demonstrate that the predicted hairstyle code $\mathbf{z}_{\text{hair}}$ and head pose ($\mathbf{z}_{\text{rot}}$ and $\mathbf{z}_{\text{tran}}$) are well separated from each other.

Finally, we anticipate an ideal model that can synthesize novel-view images for the rendered result 'Render' of a real face image 'Real'. If the model can achieve this, then naturally, it can also do the same thing to the attribute transfer result '**T.** $\mathbf{z}_*$'. As shown in the right side of Fig. 4, by changing the pose inputted into the face synthesis decoder D, we can achieve novel-view synthesis for the rendered result of a real image with 3D-aware view consistency. The above results demonstrate that our model can perform real-time and zero-shot attribute separation of a given real-face image, allowing attribute transfer and rendering at novel views without the aid of multi-view information, *i.e.*, achieving a good balance in 'zero-shot', 'attribute transfer', 'real-time' and 'realistic', as defined in Sec. 1. In Fig. 3 and Fig. 4, in order to overcome the well-known high-frequency details loss caused by training NeRF without multi-view data, we added the pre-trained DifFace as a RefineNet and demonstrated the refined results (*i.e.*, '+ RefineNet') of the rendered images after passing through the RefineNet (best viewed with zoom-in). It could be observed that the texture details of the rendered face images, such as teeth and hair, become clearer and more realistic, making our model more visually appealing while not compromising real-time performance.

Table 2: **Metric comparison of representation ability.** See the text in Sec. 4.3 for details.

Model	Image Similarity				Attribute Consistency			
	LPIPS↓	L1↓	SSIM↑	IC↑	AED↓	ATD↓	AID↓	APD↓
Deep3DFace [18]	0.3868	0.1888	0.7231	0.5779	0.0729	0.0409	0.0431	0.022
DECA [21]	0.3098	0.0948	0.5753	0.2047	0.3168	0.1636	0.2861	0.078
MoFaNeRF-fine [89]	0.3255	0.1217	0.4532	0.2118	0.6036	0.2682	0.6129	0.044
HeadNeRF [32]	0.3187	0.1260	0.7553	0.5439	0.1307	0.1137	0.3143	0.032
Ours	**0.2713**	**0.0610**	**0.8112**	**0.7151**	**0.0699**	**0.0318**	**0.0344**	**0.020**

4.3 Comparisons

Baselines. We adopt the following criteria to select baseline methods. First, the model should have the ability to perform attribute separation and novel-view rendering from a real face image, either zero-shot or by optimizing the latent codes. Second, the model and its code for obtaining the attribute parameters from real-face images should be available. The selected models include the classic regression-based explicit 3D face models Deep3DFace [18] and DECA [21], and the advanced fitting-based implicit 3D-aware face models HeadNeRF [32] and MoFaNeRF [89].

Comparison of Representation Ability. For a single-view face image, we use the facial attribute encoders of Deep3DFace, DECA, and our model to directly predict the latent attribute codes, respectively, and use the respective models to render the corresponding face image based on the obtained codes. For HeadNeRF and MoFaNeRF, we first initialize the latent attribute codes of the test image according to the methods they provide, then perform image-base fitting to obtain the optimized attribute codes, input them into the model, and render the corresponding face image, respectively. The better similarity and attribute consistency between the rendered face image and the ground truth real image could indicate a better representation ability of the latent attribute codes and a better render performance of the model.

Fig. 5 shows the prediction or fitting results for the same images ('Real'). Our results are represented by 'Ours'. MoFaNeRF also incorporates an additional pre-trained refine network to enhance the realistic texture details of the rendered face. Therefore, for a more fair comparison, we use its fitting results after refinement.

Deep3DFace is unable to represent facial areas such as hair. In some cases, there will be noticeable attribute inconsistencies in facial identity shape (row 4) and expression (rows 1-4) when compared to ground truth images. Additionally, it does not perform well in the expressiveness of some facial texture details such as beard. The face image predicted by DECA also performs poorly in terms of attribute consistency with ground truth images in terms of facial expression (rows 1-4) and texture. For example, DECA does not effectively express the texture color of the lips. In some cases, MoFaNeRF appears to completely fail to recreate a human face. It represents the facial attributes of the region excluding the hair area.

HeadNeRF achieves to fit more realistic faces with better attribute consistency in expression and texture than the above three models. However, it does not perform well in maintaining consistency in identity shape in some cases (rows 1,4), and it is unable to represent hairstyles. The face images rendered by our model demonstrate good consistency with ground truth face images in terms of facial identity shape, expression, texture, illumination, and hairstyle. This indicates that the latent attribute codes parameterized by our facial attribute encoder possess a better expressive ability for the facial attributes. We also demonstrated the refined results ('Ours-refine') of our rendered images using the refinement network DifFace as a reference, which improves the facial texture details.

Table 3: **Evaluation of attribute separation ability using Disentanglement Score (DS).** α, β, γ, η and θ stand for facial identity shape, expression, texture, illumination and head pose, respectively. See text in Sec. 4.3 for details.

Model	$DS_\alpha \uparrow$	$DS_\beta \uparrow$	$DS_\gamma \uparrow$	$DS_\eta \uparrow$	$DS_\theta \uparrow$
Deep3DFace [18]	5.76	39.6	38.3	386	**42.5**
DECA [21]	4.46	**54.8**	28.8	367	36.9
MoFaNeRF-fine [89]	3.12	21.6	21.3	-	37.8
HeadNeRF [32]	7.91	52.1	36.7	471	41.5
Ours	**8.74**	54.5	**39.1**	**476**	42.0

In Tab. 2, inspired by [23,89,32], we measure the average similarity and attribute consistency between the rendered face image and the ground truth real image with the image similarity comparison metrics: Learned Perceptual Image Patch Similarity (LPIPS) [85], L1-distance, Structural Similarity Index (SSIM), Identity Consistency (IC) and attribute consistency comparison metrics: Average Expression Distance (AED), Average Texture Distance (ATD), Average Illumination Distance (AID), and Average Pose Distance (APD). To evaluate the identity consistency (IC) between the ground truth image and the rendered face image, we compute the cosine distance of their embeddings of a pre-trained face recognition network [14]. The Average Expression Distance (AED) calculates the average 3DMM expression cosine distance between the real image and the rendered result, and the remaining three metrics of the same type, ATD, AID, and APD, are also calculated in the same manner. [18] is used to extracted the 3DMM attribute parameters. During the metric evaluation, we utilized the refined results of face images rendered by MoFaNeRF, while our model used the results without refinement.

Table 4: **Comparison of the prediction efficiency using Frames per Second (FPS).** $*$: Test on an NVIDIA RTX3090 GPU with a batch size of 1.

Model	FPS*↑
Deep3DFace [18]	**55**
DECA [21]	17
MoFaNeRF [89]	0.002
HeadNeRF [32]	0.031
Ours	15

The experimental results show that our model is capable of rendering clearer face images and performs better in terms of identity preservation. It also exhibits a good facial attribute preservation of the ground truth image in expression, texture, illumination, and head pose attributes. This indicates that our facial attribute encoder is capable of parameterizing the facial attributes of a real face to a more expressive position in latent space. Additionally, our face synthesis decoder can accurately understand the semantics of the obtained attribute codes and render the corresponding face image.

Comparison of Attribute Separation. In Tab. 3, we use the Disentanglement Score (DS) [16] to compare the models' parameterized ability in sepa-

rating facial attributes from a face image. α, β, γ, η, and θ stand for facial identity shape, expression, texture, illumination, and head pose, respectively. Ideally, when we only vary the latent code for one attribute, other facial attributes of the original rendered face image should be preserved on the re-rendered image, which was synthesized by the decoder conditioned on the modified attribute codes. We estimate the 3DMM parameters from the re-rendered image and calculate the variance of the estimated parameters (α, β, γ, η,θ). The DS_i is calculated as: $\mathrm{DS}_i = \prod_{j \neq i} \frac{var(i)}{var(j)}, i, j \in \{\alpha, \beta, \gamma, \eta, \theta\}$. A higher value of DS indicates a better separation between the specific attribute code and the remaining facial attribute codes. MoFaNeRF does not have a specific parameterization for the illumination attribute, so we did not calculate this DS_η for it. Tab. 3 shows that our model can separate facial attributes into comparatively orthogonal latent space, reducing the influence of a certain attribute code on the remaining facial attributes of the re-rendered face image.

Table 5: **Comparison of representation ability with ablated baselines.** See text in Sec. 4.4 for details.

Models	Image Similarity				Attribute Consistency			
	LPIPS↓	L1↓	SSIM↑	IC↑	AED↓	ATD↓	AID↓	APD↓
w/o $\mathcal{L}_{\mathrm{dis}}$	0.2719	0.0641	0.8106	0.7065	0.0710	0.0330	0.0353	0.020
w/o $\mathcal{L}_{\mathrm{cyc}}$	0.2854	0.0743	0.7793	0.6884	0.0750	0.0599	0.0443	0.023
w/o $\mathbf{z}_{\mathrm{hair}}$	0.2798	0.0715	0.7854	0.6954	0.0724	0.0597	0.0431	0.021
Ours	**0.2713**	**0.0610**	**0.8112**	**0.7151**	**0.0699**	**0.0318**	**0.0344**	**0.020**

Comparison of Real-time Performance. The average Frames per Second (FPS) of the different models are reported in Tab. 4. The models were tested for conducting regression-based or fitting-based prediction from a real face image on an NVIDIA RTX3090 GPU with a batch size of 1. Both MoFaNeRF and our results do not include refinement of the rendered image. The fitting-based methods, HeadNeRF and MoFaNeRF, take an average of 32 and 411 seconds, respectively, to fit and render a single image. In contrast, our regression-based model demonstrates a real-time performance that is comparable to two other regression-based models, Deep3DFace and DECA.

4.4 Ablation Study

In this section, we first attempt to verify the facial attribute encoder, the disentanglement loss, $\mathcal{L}_{\mathrm{dis}}$ (Eq. (11)), and the cycle-consistency loss, $\mathcal{L}_{\mathrm{cyc}}$ (Eq. (10)). In Tab. 5, we compare the representation ability of ablated pipelines which excludes cycle-consistency loss or disentanglement loss on the test set. The results of the complete model are comparatively better than the ablated models in the image similarity and attribute consistency between the rendered face image and the ground truth image, which implies that the face image is separated to a more expressive position in the latent space with the complete model. Tab. 5 also provides a comparison between the complete model and the ablate pipeline which excludes the hairstyle prediction head $\mathrm{PH}_{\mathrm{hair}}$ and the hairstyle attribute code $\mathbf{z}_{\mathrm{hair}}$. It shows that parameterizing facial hairstyle and projecting it into the latent space can enhance the model's ability to express facial attributes, particularly in terms of identity similarity and texture consistency.

Tab. 6 shows the intermediate results during the formal EM-like heuristic training introduced in Sec. 3.2. The 3D-aware face model has better render quality when provided with more expressive conditioned codes from the facial attribute encoder, which may be attributed to the synergy effect of the facial attribute encoder and the face synthesis decoder.

Table 6: **Comparison of representation ability during the EM-like heuristic training.**

Epoch	Image Similarity				Attribute Consistency			
	LPIPS↓	L1↓	SSIM↑	IC↑	AED↓	ATD↓	AID↓	APD↓
10	0.3244	0.1504	0.7411	0.5248	0.1801	0.1330	0.1937	0.037
50	0.2945	0.1165	0.7843	0.6063	0.1341	0.0914	0.1135	0.028
100	0.2814	0.0843	0.7994	0.6833	0.0958	0.0531	0.0704	0.023
150	0.2713	0.0610	0.8112	0.7151	0.0699	0.0318	0.0344	0.020

5 Conclusion

In this paper, we present a model that enables real-time and zero-shot attribute separation of a given real face, allowing attribute transfer and rendering at novel views without the aid of multi-view information. We achieve this by extending the alternating training algorithm to the focus that to enhance the ability of the facial attribute encoder in representing and separating attributes, and to improve the rendering quality of the 3D-aware face model. In addition, we continuously parameterize the hairstyle attribute without relying on a professional multi-view 3D scan training dataset that incorporates hair.

5.1 Limitation

Similar to [32,89,23], our model doesn't explicitly generate 3D shapes and only focuses on rendering performance. Though 3D shapes can be extracted from the neural radiance field by some means, the 3D accuracy is unwarranted. Similar to these parametric models, our model sometimes exhibits inadequate generalization in its rendering results for images that deviate significantly from the training data distribution. Besides, the training set we used does not include a dedicated multi-illumination dataset, which is inadequate for covering various types of illumination. This problem may be alleviated by searching for facial datasets that are specifically designed to capture diverse lighting conditions.

References

1. Abrevaya, V.F., Wuhrer, S., Boyer, E.: Multilinear autoencoder for 3d face model learning. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9. IEEE (2018) 4
2. Athar, S., Xu, Z., Sunkavalli, K., Shechtman, E., Shu, Z.: Rignerf: Fully controllable neural 3d portraits. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 20364–20373 (2022) 2, 4
3. Bai, Y., Fan, Y., Wang, X., Zhang, Y., Sun, J., Yuan, C., Shan, Y.: High-fidelity facial avatar reconstruction from monocular video with generative priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4541–4551 (2023) 4

4. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194 (1999) 2, 4
5. Booth, J., Roussos, A., Ponniah, A., Dunaway, D., Zafeiriou, S.: Large scale 3d morphable models. International Journal of Computer Vision **126**(2), 233–254 (2018) 4
6. Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., Dunaway, D.: A 3d morphable model learnt from 10,000 faces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5543–5552 (2016) 4
7. Bottou, L.: Stochastic gradient descent tricks. Neural Networks: Tricks of the Trade: Second Edition pp. 421–436 (2012) 13
8. Bühler, M.C., Meka, A., Li, G., Beeler, T., Hilliges, O.: Varitex: Variational neural face textures. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13890–13899 (2021) 4, 5
9. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics **20**(3), 413–425 (2013) 4
10. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022) 2, 4
11. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5799–5809 (2021) 2, 4
12. Daněček, R., Black, M.J., Bolkart, T.: Emoca: Emotion driven monocular face capture and animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20311–20322 (2022) 2, 4, 7
13. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological) **39**(1), 1–22 (1977) 3
14. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019) 9, 16
15. Deng, Y., Wang, B., Shum, H.Y.: Learning detailed radiance manifolds for high-fidelity and 3d-consistent portrait synthesis from monocular image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4423–4433 (2023) 2, 4
16. Deng, Y., Yang, J., Chen, D., Wen, F., Tong, X.: Disentangled and controllable face image generation via 3d imitative-contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5154–5163 (2020) 2, 3, 4, 5, 10, 16
17. Deng, Y., Yang, J., Xiang, J., Tong, X.: Gram: Generative radiance manifolds for 3d-aware image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10673–10683 (2022) 2, 4
18. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019) 2, 4, 5, 7, 13, 15, 16
19. Drobyshev, N., Chelishev, J., Khakhulin, T., Ivakhnenko, A., Lempitsky, V., Zakharov, E.: Megaportraits: One-shot megapixel neural head avatars. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 2663–2671 (2022) 2, 3, 4, 5
20. Egger, B., Schönborn, S., Schneider, A., Kortylewski, A., Morel-Forster, A., Blumer, C., Vetter, T.: Occlusion-aware 3d morphable models and an illumination prior for face image analysis. International Journal of Computer Vision **126**, 1269–1287 (2018) 3

21. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. vol. 40 (2021), https://doi.org/10.1145/3450626.3459936 2, 4, 5, 7, 13, 15, 16
22. Gafni, G., Thies, J., Zollhofer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8649–8658 (2021) 2, 4
23. Galanakis, S., Gecer, B., Lattas, A., Zafeiriou, S.: 3dmm-rf: Convolutional radiance fields for 3d face modeling. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3536–3547 (2023) 2, 5, 7, 16, 18
24. Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., Vetter, T.: Morphable face models-an open framework. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 75–82. IEEE (2018) 2, 4
25. Ghosh, P., Gupta, P.S., Uziel, R., Ranjan, A., Black, M.J., Bolkart, T.: Gif: Generative interpretable faces. In: 2020 International Conference on 3D Vision (3DV). pp. 868–878. IEEE (2020) 4, 5
26. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020) 2, 4
27. Grassal, P.W., Prinzler, M., Leistner, T., Rother, C., Nießner, M., Thies, J.: Neural head avatars from monocular rgb videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18653–18664 (2022) 2, 4
28. Gu, J., Liu, L., Wang, P., Theobalt, C.: Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. arXiv preprint arXiv:2110.08985 (2021) 2, 4, 8
29. Guo, X., Kan, M., Chen, T., Shan, S.: Gan with multivariate disentangling for controllable hair editing. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV. pp. 655–670. Springer (2022) 7
30. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5784–5794 (2021) 4
31. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 7
32. Hong, Y., Peng, B., Xiao, H., Liu, L., Zhang, J.: Headnerf: A real-time nerf-based parametric head model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20374–20384 (2022) 2, 3, 5, 7, 12, 13, 15, 16, 18
33. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) 11, 12
34. Kellnhofer, P., Jebe, L.C., Jones, A., Spicer, R., Pulli, K., Wetzstein, G.: Neural lumigraph rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4287–4297 (2021) 3, 5, 7
35. Khakhulin, T., Sklyarova, V., Lempitsky, V., Zakharov, E.: Realistic one-shot mesh-based head avatars. In: European Conference on Computer Vision. pp. 345–362. Springer (2022) 2, 4
36. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18750–18759 (2022) 7
37. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 13

38. Kowalski, M., Garbin, S.J., Estellers, V., Baltrušaitis, T., Johnson, M., Shotton, J.: Config: Controllable neural face image generation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 299–315. Springer (2020) 4, 5
39. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012) 9
40. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5549–5558 (2020) 12
41. Lee, Y., Choi, T., Go, H., Lee, H., Cho, S., Kim, J.: Exp-gan: 3d-aware facial image generation with expression control. In: Proceedings of the Asian Conference on Computer Vision. pp. 3812–3827 (2022) 4
42. Li, C., Morel-Forster, A., Vetter, T., Egger, B., Kortylewski, A.: Robust model-based face reconstruction through weakly-supervised outlier segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 372–381 (2023) 3
43. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017) 2, 4
44. Loper, M.M., Black, M.J.: Opendr: An approximate differentiable renderer. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13. pp. 154–169. Springer (2014) 4
45. Ma, Z., Zhu, X., Qi, G.J., Lei, Z., Zhang, L.: Otavatar: One-shot talking face avatar with controllable tri-plane rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16901–16910 (2023) 2, 4
46. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4460–4470 (2019) 4
47. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 4
48. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017) 12
49. Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., Yang, Y.L.: Hologan: Unsupervised learning of 3d representations from natural images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7588–7597 (2019) 4, 5
50. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2021) 4
51. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3504–3515 (2020) 4
52. Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I.: Stylesdf: High-resolution 3d-consistent image and geometry generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13503–13513 (2022) 2, 4
53. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 165–174 (2019) 4
54. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5865–5874 (2021) 2, 4

55. Piao, J., Sun, K., Wang, Q., Lin, K.Y., Li, H.: Inverting generative adversarial renderer for face reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15619–15628 (2021) 4, 5
56. Ramamoorthi, R., Hanrahan, P.: An efficient representation for irradiance environment maps. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 497–500 (2001) 6
57. Ramon, E., Triginer, G., Escur, J., Pumarola, A., Garcia, J., Giro-i Nieto, X., Moreno-Noguer, F.: H3d-net: Few-shot high-fidelity 3d head reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5620–5629 (2021) 4
58. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3d faces using convolutional mesh autoencoders. In: Proceedings of the European conference on computer vision (ECCV). pp. 704–720 (2018) 4
59. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d. arXiv preprint arXiv:2007.08501 (2020) 4
60. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13759–13768 (2021) 2, 4
61. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. ACM Transactions on graphics (TOG) **42**(1), 1–13 (2022) 4
62. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: Generative radiance fields for 3d-aware image synthesis. Advances in Neural Information Processing Systems **33**, 20154–20166 (2020) 4
63. Shoshan, A., Bhonker, N., Kviatkovsky, I., Medioni, G.: Gan-control: Explicitly controllable gans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14083–14093 (2021) 2, 3, 4, 5
64. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in neural information processing systems **32** (2019) 2, 4
65. Siarohin, A., Menapace, W., Skorokhodov, I., Olszewski, K., Ren, J., Lee, H.Y., Chai, M., Tulyakov, S.: Unsupervised volumetric animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4658–4669 (2023) 4
66. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 9
67. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. Advances in neural information processing systems **33**, 7462–7473 (2020) 4
68. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. Advances in Neural Information Processing Systems **32** (2019) 4
69. Sun, K., Wu, S., Huang, Z., Zhang, N., Wang, Q., Li, H.: Controllable 3d face synthesis with conditional generative occupancy fields. arXiv preprint arXiv:2206.08361 (2022) 2, 4, 5
70. Tewari, A., Elgharib, M., Bernard, F., Seidel, H.P., Pérez, P., Zollhöfer, M., Theobalt, C.: Pie: Portrait image embedding for semantic control. ACM Transactions on Graphics (TOG) **39**(6), 1–14 (2020) 5, 10
71. Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H.P., Pérez, P., Zollhofer, M., Theobalt, C.: Stylerig: Rigging stylegan for 3d control over portrait images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6142–6151 (2020) 2, 3, 4, 5, 10
72. Tewari, A., Pan, X., Fried, O., Agrawala, M., Theobalt, C., et al.: Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1516–1525 (2022) 4, 5

73. Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 1274–1283 (2017) 13
74. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7346–7355 (2018) 2, 7
75. Wang, X., Guo, Y., Yang, Z., Zhang, J.: Prior-guided multi-view 3d head reconstruction. IEEE Transactions on Multimedia **24**, 4028–4040 (2021) 4
76. Wu, S., Yan, Y., Li, Y., Cheng, Y., Zhu, W., Gao, K., Li, X., Zhai, G.: Ganhead: Towards generative animatable neural head avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 437–447 (2023) 2, 5, 7
77. Wu, Y., Deng, Y., Yang, J., Wei, F., Chen, Q., Tong, X.: Anifacegan: Animatable 3d-aware face image generation for video avatars. arXiv preprint arXiv:2210.06465 (2022) 2, 4, 5
78. Xu, S., Yang, J., Chen, D., Wen, F., Deng, Y., Jia, Y., Tong, X.: Deep 3d portrait from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7710–7720 (2020) 12
79. Yao, S., Zhong, R., Yan, Y., Zhai, G., Yang, X.: Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. arXiv preprint arXiv:2201.00791 (2022) 4
80. Yenamandra, T., Tewari, A., Bernard, F., Seidel, H., Elgharib, M., Cremers, D., Theobalt, C.: i3dmm: Deep implicit 3d morphable model of human heads. In: Proceedings of the IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2021) 5
81. Yenamandra, T., Tewari, A., Bernard, F., Seidel, H.P., Elgharib, M., Cremers, D., Theobalt, C.: i3dmm: Deep implicit 3d morphable model of human heads. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12803–12813 (2021) 2, 5, 7
82. Yin, Y., Ghasedi, K., Wu, H., Yang, J., Tong, X., Fu, Y.: Nerfinvertor: High fidelity nerf-gan inversion for single-shot real image animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8539–8548 (2023) 4
83. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 325–341 (2018) 12
84. Yue, Z., Loy, C.C.: Diface: Blind face restoration with diffused error contraction. arXiv preprint arXiv:2212.06512 (2022) 3, 12
85. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 16
86. Zheng, M., Yang, H., Huang, D., Chen, L.: Imface: A nonlinear 3d morphable face model with implicit neural representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20343–20352 (2022) 2, 4
87. Zheng, Y., Abrevaya, V.F., Bühler, M.C., Chen, X., Black, M.J., Hilliges, O.: Im avatar: Implicit morphable head avatars from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13545–13555 (2022) 2, 4
88. Zheng, Y., Yifan, W., Wetzstein, G., Black, M.J., Hilliges, O.: Pointavatar: Deformable point-based head avatars from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21057–21067 (2023) 4
89. Zhuang, Y., Zhu, H., Sun, X., Cao, X.: Mofanerf: Morphable facial neural radiance field. In: European Conference on Computer Vision (2022) 2, 5, 7, 12, 15, 16, 18
90. Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4574–4584 (2023) 4