# Learning to Generate Conditional Tri-plane for 3D-aware Expression Controllable Portrait Animation

Taekyung Ki<sup>1</sup><sup>®</sup>, Dongchan Min<sup>2</sup><sup>®</sup>, and Gyeongsu Chae<sup>1</sup><sup>®</sup>

<sup>1</sup> DeepBrainAI Inc., South Korea <sup>2</sup> Graduate School of AI, KAIST, South Korea taek@deepbrain.io alsehdcks95@kaist.ac.kr gc@deepbrain.io https://deepbrainai-research.github.io/export3d

Abstract. In this paper, we present Export3D, a one-shot 3D-aware portrait animation method that is able to control the facial expression and camera view of a given portrait image. To achieve this, we introduce a tri-plane generator with an effective expression conditioning method, which directly generates a tri-plane of 3D prior by transferring the expression parameter of 3DMM into the source image. The tri-plane is then decoded into the image of different view through a differentiable volume rendering. Existing portrait animation methods heavily rely on image warping to transfer the expression in the motion space, challenging on disentanglement of appearance and expression. In contrast, we propose a contrastive pre-training framework for appearance-free expression parameter, eliminating undesirable appearance swap when transferring a cross-identity expression. Extensive experiments show that our pre-training framework can learn the appearance-free expression representation hidden in 3DMM, and our model can generate 3D-aware expression controllable portrait images without appearance swap in the cross-identity manner.

Keywords: Portrait Image Animation  $\cdot$  Facial Expression Control  $\cdot$  3D-aware Synthesis

# 1 Introduction

Portrait image animation aims to generate a video of a given source identity with the driving motion. It has received a lot of attention due to the potential of virtual human services, such as cross-lingual film dubbing [12, 28], virtual avatar chatting [37, 64], and video conferencing [51, 53]. In these scenarios, it is essential to transfer the facial expression (e.g., eye-blinking, lip motion, etc.) from different person, i.e., *cross-identity transfer*, while preserving the source identity. However, it is challenging due to the ambiguity between appearance and expression [17] and the lack of paired data (e.g., different faces with the same expression) for disentanglement representation learning [38].

Most 2D-based methods rely on image warping [23,48,54,60,65], which warps the source image to the driving image by estimating the motion between them. To impose a bottleneck for the motion representation, they encode the motion into the difference between sparse key-points [23, 48, 65] or latent codes [54], which are trained in an unsupervised manner. However, in this scenario, the facial expressions are encoded into the motion space as well, in terms of local motion, which tends to be neglected due to the relatively large head motions. Furthermore, since the facial expression and the appearance are highly entangled in the image space, cross-identity expression transfer often involves the source appearance change. DPE [38] tackles this entanglement issue by proposing a self-supervised disentanglement learning framework based on cycle-consistency learning [66]. However, it shows temporal inconsistency in the generated video due to its instability of cycle-consistency learning.

Another line of works [30, 31, 34, 62] explores facial expression control in 3D space using the neural radiance fields (NeRFs) [36]. They leverage pre-trained latent representation of 3D GAN [9] for 3D facial prior where they design the expression in terms of latent code [34, 62] or predict deformation field [39] to deform the well-constructed 3D representation, such as tri-plane [30, 31, 34, 62]. However, the latent code cannot faithfully reconstruct the source identity [34], and the point-wise deformation fields to those 3D representations yield video-level artifacts, such as flickers [30].

In this paper, we address the appearance-expression entanglement issue by proposing a contrastive pre-training framework over video datasets that produces appearance-free facial expressions with an orthogonal structure. Armed with this representation, we build a one-shot 3D-aware portrait image animation method, namely Export3D, which controls the facial expression and 3D camera view of a given source image without appearance swap. To achieve this, we design a generator architecture consisting of vision transformer (ViT) and convolution layers [16, 40, 51] that directly generates the tri-planes from the source image and driving expression parameters. Instead of predicting the deformation fields for the expression, we introduce an expression adaptive layer normalization (EAdaLN) which can effectively transfer the driving expression to the source image. The main contributions of this work are summarized as follows:

- We present Export3D, a one-shot 3D-aware portrait image animation method that can explicitly control the facial expression and camera view of the source image only using the expression and camera parameters.
- We propose a contrastive pre-training framework for the appearancefree facial expression distilled from the 3DMM parameters where they form an orthogonal structure for different facial expressions.
- Extensive experiments demonstrate that our pre-training framework can learn the appearance-free expression, which enables our method to transfer the cross-identity expression without undesirable appearance swap.

### 2 Related Works

#### 2.1 3D-aware Image Synthesis

3D-aware image synthesis aims to generate images with explicit camera pose control [9, 10, 14, 20, 46, 56]. This is achieved by conditioning the camera pose parameter into generative features, which are then rendered into an RGB image through differentiable volume rendering [6, 32, 36, 39]. This rendering technique has integrated with adversarial learning [9, 10, 14, 19, 20, 46, 56] to learn 3D view consistency from the unposed dataset. GRAM [14] generates a multi-view consistent image by learning the radiance field on a set of 2D surface manifolds. AniFaceGAN [55] further learns the deformation fields [39] for the facial expression on these manifolds [14] for explicit facial expression control. EG3D [9] introduces a tri-plane representation that provides a strong 3D position encoding with neural volume rendering and become the one of the most prominent representation in this field. However, these methods generate portrait images from noise, requiring further process for real image manipulation.

Relying on the generateive power of EG3D, several works [7, 29, 34, 49, 51, 58, 61–63] extend 2D GAN-inversion [1, 2, 44, 50] methods, which is challenging due to the multi-view consistency for a single-view image. Specifically, based on facial symmetry, SPI [61] utilizes horizontally flipped images for pseudo supervision to the occluded facial region. However, it requires multi-stage latent code optimizations. GOAE [63] proposes an encoder-based inversion for EG3D which enhances multi-view consistency via an occlusion-aware tri-plane mixing module. Live3DPortrait [51] can reconstruct multi-view consistent portrait images by leveraging the synthetic data of pre-trained EG3D to provide multi-view supervision. However, these methods cannot explicitly manipulate the expression of the source image.

We propose a tri-plane generator architecture that can generate the tri-plane of a given source image with explicit expression control. Inspired by [40,51], we design this generator with ViT and convolution layers [16], and directly inject expression parameters into the tri-plane generating process through the expression adaptive layer normalization (EAdaLN). By leveraging the strong power of NeRF [9,36,49,51,62], we decode the generated tri-plane into multi-view images with explicit expression manipulation.

#### 2.2 Portrait Image Animation

Portrait image animation, or face reenactment, is a task that animates a given source image according to the input driving condition, either audio [21,28,33,37, 41,64] or image [48,53,54,60,65]. Specifically, image-driven methods transfer the motion of the driving image into the source image by learning the motion between them. Most works [48, 53, 65] use facial key-points as a pivot representation to be aware of motion via the key-point displacement. FOMM [48] estimates facial key-points in an unsupervised manner, approximating the motion through the first-order Taylor expansion. LIA [54] encodes a motion in terms of latent



Fig. 1: Training overview of Export3D. We convert a source image  $S \in \mathbb{R}^{3 \times H \times W}$ into a tri-plane  $T_{\beta_D}(S)$  for rich 3D priors, conditioned on an expression parameter  $\beta_D \in \mathbb{R}^{64}$  from a driving image  $D \in \mathbb{R}^{3 \times H \times W}$ . A differentiable volume rendering renders the tri-plane into a raw rendered image  $\hat{D}_{raw} \in \mathbb{R}^{3 \times \frac{H}{4} \times \frac{W}{4}}$  using the camera parameter  $p_D \in \mathbb{R}^{25}$  of D, which is then super-resolved into a final image  $\hat{D} \in \mathbb{R}^{3 \times H \times W}$ .

codes by introducing an orthonormal basis as a motion dictionary. However, the local motion (e.g., facial expression) and the global motion (e.g., head motion) are still entangled in those representations. DPE [38] proposes a bidirectional cyclic training strategy to decouple the pose and expression within the latent codes, while it produces video-level artifacts due to the instability of the cycle-consistency learning.

To explicitly control the facial expression, several works leverage the expression parameters of 3D morphable models (3DMM) [8] in 2D [18,60] or 3D spaces [30,31,34]. StyleHEAT [60] uses 3DMM to warp 2D spatial features of pre-trained StyleGAN2 [25] while yielding texture sticking. OTAvatar [34] proposes a one-shot test-time optimization method that optimizes identity codes of a single source image and learns expression-aware motion latent codes in the latent space of pre-trained EG3D. HiDe-NeRF [30] and NOFA [62] take a different way by predicting an expression-aware deformation field [39] that deforms the tri-plane [9] reconstructed from the source image.

Our method belongs to image-driven approaches, distinguishing itself by not depending on 2D image warping or 3D deformation fields. Toward this, we propose the generator architecture that uses a source image and driving expression parameters to produce an expression-transferred tri-plane, wherein the expression parameters directly modulate the source visual features through the expression adaptive layer normalization (EAdaLN). Furthermore, we mitigate the appearance swap issue inherent in transferring other person's expression by introducing a contrastive pre-training method to obtain appearance-free expression representations.

### 3 Methods

First of all, we formulate our portrait animation method, *Export3D*. Given a source image  $S \in \mathbb{R}^{3 \times H \times W}$ , our method transfers the facial expression and camera view of a driving image  $D \in \mathbb{R}^{3 \times H \times W}$  with the expression and camera parameters, respectively. We employ a tri-plane [9] as the intermediate feature representation, providing a strong 3D position information for differentiable volume rendering [32,36]. We directly generate an expression-transferred tri-plane from the source image and the driving expression parameter [3,8] through expression adaptive layer normalization (EAdaLN) (Sec. 3.2). Based on the observation that the expression parameter still contains the appearance information, we propose a pre-training framework using contrastive learning to obtain the appearance-free expression, which forms an orthogonal structure for different expressions (Sec. 3.1). The expression-transferred tri-plane is rendered into a 3D-aware image through the differentiable volume rendering, and then super-resolved into the final output (Sec. 3.3).

#### 3.1 Contrastive Learned Basis Scaling (CLeBS)

Natural speaking style comes from the the non-verbal component, such as eye blinking. To explicitly control the expression of the generated face, we utilize the expression parameter  $\beta \in \mathbb{R}^{64}$  from the widely used 3D morphable models (3DMM) [8] in 3D face reconstruction. However, simply using those parameters for transferring the other person's expression fails to preserve the facial identity of the source face.

**Disentangling Expression and Appearance**. In 3DMM-based face reconstruction, the identity-appearance has been rarely explored. However, [17] shows that a 3D face shape can be reconstructed only using the expression parameters not using the shape parameters, or vice versa. We also observe



Fig. 2: Contrastive pre-training framework for LeBS. We sample the positive and the negative samples from the same video source so that those samples share the same appearance. Using contrastive learning, the encoder  $f_e(\cdot)$  learns an appearance-free representations.

that the expression parameter of 3DMM is highly entangled with the appearance (Fig. 8a), resulting in an undesirable appearance swap when transferring the cross-identity expressions. We assume that the expression parameter needs to be refined to represent pure facial expressions. To address this issue, we propose a contrastive learning based pre-training framework [11, 22, 37, 42] on video dataset to discard the appearance information hidden in the expression parameter. Specifically, given a video sequence  $\{X_i\}_{i=1}^N$  and its corresponding expression sequence  $\{\beta_i\}_{i=1}^N$ , we sample an aligned image-expression pair  $(X_k, \beta_k)$  for the

positive and the non-aligned pairs for the negatives as illustrated in Fig. 2. The images and the expressions are mapped into *d*-dimensional representations, and the distance between the positive (or negative) representation pairs is minimized (or maximized) via the following contrastive objective  $\mathcal{L}_{cl}$ :

$$\mathcal{L}_{cl} = -\log\left(\frac{\exp(\cos(f_I(X_k), f_e(\beta_k))/\tau)}{\sum_{j \neq k} \exp(\cos(f_I(X_j), f_e(\beta_k))/\tau)}\right),\tag{1}$$

where  $f_I(\cdot)$  is an image encoder,  $f_e(\cdot)$  is an expression encoder,  $\tau$  is the temperature, and  $\cos(\cdot, \cdot)$  is the cosine similarity, respectively. Since all samples are from the same video, they share the same appearance, thereby Eq. (1) enforces the encoders to learn appearance-free expression.

Moreover, for designing the expression encoder  $f_e(\cdot)$ , we focus on the orthogonal structure of 3DMM [8] that controls different expressions along different orthogonal directions. To provide the appearance-free expression with the orthogonal structure, we introduce a learned orthonormal basis V:

$$V = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\} \subseteq \mathbb{R}^d \quad \text{and} \quad \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \delta_{ij}, \ \forall i, j, \tag{2}$$

spanning our new expression sub-space ( $\delta_{ij}$  is the Kroneker delta function). More precisely, we convert the expression  $\beta \in \mathbb{R}^{64}$  into the low-dimensional coefficient  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n) \in \mathbb{R}^n$  ( $n \ll 64$ ) and then scales the learned orthonormal basis  $V \subseteq \mathbb{R}^d$  to produce the appearance-free expression representation  $\beta' \in \mathbb{R}^d$ :

$$\beta' = f_e(\beta) = \lambda_1 \mathbf{v}_1 + \lambda_2 \mathbf{v}_2 + \dots + \lambda_n \mathbf{v}_n \in \mathbb{R}^d.$$
(3)

We apply QR-decomposition [54] to a learned weight  $(\in \mathbb{R}^{d \times n})$  to explicitly compute the orthonormal basis  $V \in \mathbb{R}^{d \times n}$ . In this space, an expression is a linear combination of the basis  $V = \{\mathbf{v}_i\}_{i=1}^n$  where the coefficient  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$ is responsible for the intensity of each expression direction. We call our encoder  $f_e(\cdot)$  a learned basis scaling (LeBS) module with contrastive pre-training (CLeBS). Once CLeBS is pre-trained with Eq. (1), no further training is required as illustrated in Fig. 1 and Fig. 3, and the image encoder  $f_I(\cdot)$  is never used after then.

#### 3.2 Hybrid Tri-plane Generator

We employ the tri-plane as the intermediate feature representation for 3D prior to volume rendering. Tri-plane T consists of features assigned on the 3 axisaligned planes (i.e., xy, yz, zx planes):

$$\mathbf{T} = (\mathbf{T}_{xy}, \mathbf{T}_{yz}, \mathbf{T}_{zx}) \in \mathbb{R}^{3 \times 32 \times \frac{H}{2} \times \frac{W}{2}},\tag{4}$$

where  $T_{ij} \in \mathbb{R}^{32 \times \frac{H}{2} \times \frac{W}{2}}$  is the 32-dimensional feature of  $\frac{H}{2} \times \frac{W}{2}$  resolution on the *ij*-plane. EG3D [9] utilizes StyleGAN2 [25] to generate the tri-plane from a noise, forming the style latent space  $\mathcal{W} \subseteq \mathbb{R}^{512}$ . Several works [7, 29, 34, 49, 58, 61, 62] extend the 2D GAN-inversion methods [1, 2, 44, 45, 50] to 3D GANinversion in terms of reconstructing the tri-plane from the style latent code.

However, these methods often face challenges in preserving facial identity since the style latent code lacks the capacity for encoding spatial information and person-specific visual details. We directly generate the expressiontransferred tri-plane  $T_{\beta_D}(S)$  from the source S and the driving expression  $\beta_D \in \mathbb{R}^{64}$  to reconstruct the driving D. Inspired by Live3DPortrait [51], we construct the tri-plane generator with ViT and convolution [16, 57]. Specifically, we convert  $S \in$  $\mathbb{R}^{3 \times H \times W}$  into a visual feature in  $\mathbb{R}^{\frac{h}{4} \times \frac{H}{2^3} \times \frac{W}{2^3}}$  through a stack of convolutional blocks, and then merge it into the *h*-dimensional  $\frac{H \cdot W}{2^8}$  visual tokens through a overlap patch merge operator [57]. These tokens and driving expression are processed through a conditional ViT [16, 40, 57] blocks, namely EAdaLN-ViT, where the expression modulates [40] the visual tokens through expression adaptive layer normalization (EAdaLN) as il-



Fig. 3: Hybrid tri-plane generator G and Expression Adaptive Layer Normalization (EAdaLN). EAdaLN modulates the expression of S using the refined expression  $\beta'$  from CLeBS.

lustrated in Fig. 3. EAdaLN is applied right before the multi-head self-attention (MSA) and the mix feed-forward network (Mix-FFN) [57] of each ViT block to inject the semantic expression into the visual tokens:

$$\mathrm{EAdaLN}(x,\beta'_D) = \sigma(\beta'_D) \times \mathrm{LN}(x) + \mu(\beta'_D) \in \mathbb{R}^{h \times \left(\frac{H \cdot W}{2^8}\right)},\tag{5}$$

where x is the input visual token,  $LN(\cdot)$  is the layer normalization [5],  $\sigma(\beta'_D)$ and  $\mu(\beta'_D)$  are the h-dimensional scale and shift factors computed from  $\beta'_D = f_e(\beta_D) \in \mathbb{R}^d$ , respectively. To efficiently propagate the visual tokens to the higher resolution, we upsample the visual tokens with pixel shuffle [51] followed by the Gaussian low-pass filter [24]. We experimentally find that the tokens and the pixel shuffle produce grid artifacts, challenging to eliminate in the image space. Employing low-pass filters effectively mitigates these artifacts by smoothing the borderline artifacts over the coordinate. Lastly, we use ViT and convolutional blocks to output the tri-plane  $T_{\beta_D}(S)$ :

$$T_{\beta_D}(S) = \mathbf{G}(S, \beta_D) \in \mathbb{R}^{3 \times 32 \times \frac{H}{2} \times \frac{W}{2}}.$$
(6)

Note that our method does not query the expression parameter to estimate the motion [18, 38, 60], rather it is used as the multi-dimensional label. To stabilize the tri-plane generation, we incorporate the online exponential moving average

(EMA) over tri-plane  $T_{EMA}$  which is added to the generated tri-plane. Please refer to supplementary materials for detailed architectures.

#### 3.3 Volume Rendering and Super-resolution

The tri-plane can be rendered into a 2D RGB image through the differentiable volume rendering [9,32,36]. The expression-transferred tri-plane  $T_{\beta_D}(S)$  is projected onto 3 orthogonal planes (xy, yz, zx-planes) and then aggregated through average [9]:

$$F_{\beta_D}(S) = \frac{1}{3} (F_{\beta_D, xy}(S) + F_{\beta_D, yz}(S) + F_{\beta_D, zx}(S)),$$
(7)

where  $F_{\beta_D,ij}(S)$  are the projected features of  $T_{\beta_D}(S)$  onto the *ij* planes. A lightweight MLP assigns a color *c* and density  $\sigma$  to each point (x, y, z) using the aggregated feature  $F_{\beta_D}(S)$ :

$$MLP: F_{\beta_D}(S) \longrightarrow (c, \sigma).$$
(8)

The differentiable volume rendering [9,21,36] composites each color c and density  $\sigma$  into a RGB value C along the camera ray **r**:

$$\mathcal{C} = \int_{t_n}^{t_f} \sigma(\mathbf{r}(t)) \cdot c(\mathbf{r}(t)) \cdot T(t) dt, \qquad (9)$$

where  $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}, t \in [t_n, t_f]$ , with camera center  $\mathbf{o} \in \mathbb{R}^3$ , viewing direction  $\mathbf{d} \in \mathbb{R}^3$ , and T(t) is the accumulation measure along the ray  $\mathbf{r}$  from  $t_n$  to t:

$$T(t) = \exp\left(-\int_{t_n}^t \sigma(\mathbf{r}(s))ds\right).$$
 (10)

Note that the ray **r** is determined by the driving camera parameter  $p_D \in \mathbb{R}^{25}$  to render the generated tri-plane  $T_{\beta_D}(S)$  into a image of the same view with D. As the appearance and the expression are already encoded in the tri-plane generation, the volume rendering can determine the view-consistent images.

Directly rendering a target high-resolution image requires high computational cost. One promising approach to address this issue is to incorporate superresolution blocks [9,51,56] that upsamples the rendered image of low resolution. Following this approach, we first render a  $\hat{D}_{raw} \in \mathbb{R}^{3 \times \frac{H}{4} \times \frac{W}{4}}$  and then apply the super-resolution to obtain the target resolution  $\hat{D} \in \mathbb{R}^{3 \times H \times W}$ , as illustrated in Fig. 1. Instead of using style-modulated convolution [9,51], we use plane convolutional blocks for super-resolution, as we do not leverage the style latent code. Detailed architecture is provided in supplementary materials.

### 4 Experiments

#### 4.1 Dataset and Pre-processing

We train our model on real video dataset VFHQ [59]. Following the video preprocessing strategies in [28,48], we convert the original video into 25 fps and crop

Table 1: Quantitative comparison on VFHQ. The best score for each metric is in **bold**. Note that we only measure CSIM [13], AED and APD [15, 43] for the crossidentity experiment as no ground-truth is available. <sup>†</sup>: Evaluated only on the foreground facial region.

0.811

0.082

0.030

0.694

0.208

| Madal                       | Same-identity           |                         |                          |                         |                          |                          |                         | Cross-identity           |                          |  |
|-----------------------------|-------------------------|-------------------------|--------------------------|-------------------------|--------------------------|--------------------------|-------------------------|--------------------------|--------------------------|--|
| Model                       | $\mathbf{PSNR}\uparrow$ | $\mathbf{SSIM}\uparrow$ | $\mathbf{AKD}\downarrow$ | $\mathbf{CSIM}\uparrow$ | $\mathbf{AED}\downarrow$ | $\mathbf{APD}\downarrow$ | $\mathbf{CSIM}\uparrow$ | $\mathbf{AED}\downarrow$ | $\mathbf{APD}\downarrow$ |  |
| StyleHEAT [60]              | 14.233                  | 0.428                   | 30.406                   | 0.464                   | 0.161                    | 0.139                    | 0.505                   | 0.242                    | 0.136                    |  |
| DPE [38]                    | 23.241                  | 0.750                   | 3.661                    | 0.831                   | 0.083                    | 0.032                    | 0.586                   | 0.253                    | 0.085                    |  |
| $ROME^{\dagger}$ [27]       | 14.185                  | 0.642                   | 7.281                    | 0.737                   | 0.111                    | 0.051                    | 0.641                   | 0.224                    | 0.074                    |  |
| OTAvatar <sup>†</sup> [34]  | 17.441                  | 0.651                   | 11.502                   | 0.662                   | 0.176                    | 0.067                    | 0.610                   | 0.290                    | 0.198                    |  |
| HiDe-NeRF <sup>†</sup> [30] | 21.228                  | 0.728                   | 8.245                    | 0.867                   | 0.106                    | 0.049                    | 0.707                   | 0.255                    | 0.065                    |  |

3.453



Fig. 4: Comparison on same-identity experiments. For a fair comparison, we follow the pre-processing strategy of each method.

the facial regions of resolution  $256 \times 256$ , ensuring that the nose is located at the center of the image. We use a 3DMM extractor [15] to obtain the expression parameter  $\beta \in \mathbb{R}^{64}$ . We adopt the pre-preprocessing strategy of EG3D [9] for the camera parameter  $p \in \mathbb{R}^{25}$  (the concatenation of the camera intrinsic parameters in  $\mathbb{R}^9$  and the inverse extrinsic parameter in  $\mathbb{R}^{16}$ ). After the video pre-processing, 6196 video clips are used for training, and 50 videos are used for test. We also evaluate our model on the test dataset of TalkingHead-1KH [53]. After the same pre-processing, remaining 20 videos of different identities are used.

#### 4.2Evaluation

Ours

23.555

0.704

We compare our model against 2D-based [38,60] and 3D-based [27,30,34] imagedriven portrait animation methods whose official implementations are available. StyleHEAT [60] warps the 2D spatial features of pre-trained StyleGAN2 using

0.080

Table 2: Quantitative comparison on TalkingHead-1KH. The best score for each metric is in **bold**. Note that we only measure CSIM [13], AED and APD [15,43] for the cross-identity experiment as no ground-truth is available.

| : | Evaluated | only c | on the | foreground | facial | region. |
|---|-----------|--------|--------|------------|--------|---------|
|---|-----------|--------|--------|------------|--------|---------|

| Mathad                      |                         |                         | Same-id                  | lentity                 |                          |                          | Cross-identity          |                          |                          |  |
|-----------------------------|-------------------------|-------------------------|--------------------------|-------------------------|--------------------------|--------------------------|-------------------------|--------------------------|--------------------------|--|
| Method                      | $\mathbf{PSNR}\uparrow$ | $\mathbf{SSIM}\uparrow$ | $\mathbf{AKD}\downarrow$ | $\mathbf{CSIM}\uparrow$ | $\mathbf{AED}\downarrow$ | $\mathbf{APD}\downarrow$ | $\mathbf{CSIM}\uparrow$ | $\mathbf{AED}\downarrow$ | $\mathbf{APD}\downarrow$ |  |
| StyleHEAT [60]              | 15.613                  | 0.517                   | 21.198                   | 0.575                   | 0.148                    | 0.095                    | 0.571                   | 0.218                    | 0.102                    |  |
| DPE [38]                    | 23.201                  | 0.786                   | 4.281                    | 0.807                   | 0.093                    | 0.029                    | 0.714                   | 0.216                    | 0.081                    |  |
| $ROME^{\dagger}$ [27]       | 15.921                  | 0.695                   | 13.444                   | 0.726                   | 0.123                    | 0.062                    | 0.667                   | 0.201                    | 0.084                    |  |
| OTAvatar <sup>†</sup> [34]  | 16.952                  | 0.660                   | 11.615                   | 0.668                   | 0.181                    | 0.063                    | 0.682                   | 0.247                    | 0.150                    |  |
| HiDe-NeRF <sup>†</sup> [30] | 19.759                  | 0.729                   | 5.746                    | 0.843                   | 0.112                    | 0.043                    | 0.757                   | 0.232                    | 0.085                    |  |
| Ours                        | 23.239                  | 0.797                   | 3.581                    | 0.764                   | 0.092                    | 0.033                    | 0.772                   | 0.204                    | 0.076                    |  |



Fig. 5: Comparison on cross-identity experiments. For a fair comparison, we follow the pre-processing strategy of each method. Notably, most portrait animation methods fail to preserve the source identity or transfer driving appearance features, such as *eye shape and facial contour*, in cross-identity scenarios.

3DMM parameters, **DPE** [38] disentangles the pose and the expression in the motion latent space without using 3DMM parameters. **ROME** [27] is a meshbased method transferring the expression and pose using 3DMM. **OTAvatar** [34] leverages pre-trained EG3D [9] by modeling head motion in terms of latent codes. **HiDe-NeRF** [30] deforms the source tri-plane by predicting expressionaware deformation fields. For evaluation, we employ peak signal-to-noise ratio (**PSNR**) and structural similarity index measure (**SSIM**) for image quality, average key-point distance (**AKD**) [43] for facial structure based on the 68 facial key-points, cosine similarity of identity embedding (**CSIM**) [13] for identity preservation, average expression distance (**AED**), and average pose distance (**APD**) [15, 43] for expression transferring and pose matching. For the cross-

Table 3: Ablation studies on the expression encoding. Same evaluation setting with Tab. 1. The best score for each metric is in **bold**.

| Mathad              |                         |                         | Same-ic                    | lentity                 | Cro                        | Cross-identity             |                         |                            |                            |
|---------------------|-------------------------|-------------------------|----------------------------|-------------------------|----------------------------|----------------------------|-------------------------|----------------------------|----------------------------|
| Method              | $\mathbf{PSNR}\uparrow$ | $\mathbf{SSIM}\uparrow$ | $\mathbf{AKD}{\downarrow}$ | $\mathbf{CSIM}\uparrow$ | $\mathbf{AED}{\downarrow}$ | $\mathbf{APD}{\downarrow}$ | $\mathbf{CSIM}\uparrow$ | $\mathbf{AED}{\downarrow}$ | $\mathbf{APD}{\downarrow}$ |
| Direct 3DMM         | 23.077                  | 0.688                   | 3.874                      | 0.789                   | 0.105                      | 0.044                      | 0.648                   | 0.209                      | 0.073                      |
| E2E LeBS $(n = 25)$ | 23.105                  | 0.672                   | 3.775                      | 0.745                   | 0.109                      | 0.040                      | 0.670                   | 0.218                      | 0.071                      |
| E2E LeBS $(n = 10)$ | 23.235                  | 0.676                   | 3.755                      | 0.751                   | 0.110                      | 0.038                      | 0.672                   | 0.238                      | 0.079                      |
| E2E LeBS $(n = 5)$  | 22.631                  | 0.646                   | 4.114                      | 0.658                   | 0.140                      | 0.046                      | 0.632                   | 0.246                      | 0.076                      |
| Full (CLeBS)        | 23.555                  | 0.704                   | 3.453                      | 0.811                   | 0.082                      | 0.030                      | 0.694                   | 0.208                      | 0.080                      |



Fig. 6: Ablation studies on the expression encoding. Without our contrastive pre-training, the expression encoders transfer the expression together with the appearance, such as *eyelids and the head size*.

identity experiments, we only measure CSIM, AED and APD as no ground-truth image is available.

Same-identity experiments. We report the same-identity transfer experiment results in Tab. 1 and Tab. 2, and illustrate the qualitative results in Fig. 4. For a fair comparison, ROME [27], OTAvatar [34], and HiDe-NeRF [34] are evaluated on the foreground facial region with different field of view. DPE [38] shows the stable performance in the same-identity experiments with the fine-grained expression controls. Among the 3D-based methods, HiDe-NeRF [30] scores the highest in the identity preservation (CSIM). Our method scores the best result in the majority of evaluation metrics. Especially, it has an advantage in expression controls (AKD and AED).

**Cross-identity experiments.** In Tab. 1 and Tab. 2, we also conduct the cross-identity transfer experiments that transfers the expression and pose of different identity into the source identity. As illustrated in Fig. 5, DPE [38] shows visual artifacts and appearance swap, such as face contours and eye shape, due to the insufficient disentanglement of expression and pose in the motion space. HiDe-NeRF [30] scores the highest identity preservation (CSIM) while un-predictable light changes are involved due to the point-wise deformation field on the tri-plane. Our method can transfer the driving expression without ap-



Fig. 7: Cross-attention vs. EAdaLN.

| Table                | 4:     | Ablation    | s     | tudies    | on     |
|----------------------|--------|-------------|-------|-----------|--------|
| EAdaL                | N. Th  | e best sco  | re fo | r each n  | netric |
| is in $\mathbf{bol}$ | d. We  | replace E   | Adal  | LN in $G$ | with   |
| cross-att            | entior | n to verify | the   | effectiv  | reness |
| of EAda              | LN.    |             |       |           |        |

| Mathod                    | Sa                      | ne-identi                 | ty              | Cross-identity |                          |                 |  |
|---------------------------|-------------------------|---------------------------|-----------------|----------------|--------------------------|-----------------|--|
| Method                    | $\mathbf{CSIM}\uparrow$ | $\mathbf{AED} \downarrow$ | $APD\downarrow$ | $CSIM\uparrow$ | $\mathbf{AED}\downarrow$ | $APD\downarrow$ |  |
| Ours (w. Cross-attention) | 0.678                   | 0.125                     | 0.042           | 0.631          | 0.271                    | 0.122           |  |
| Ours (w. EAdaLN)          | 0.811                   | 0.082                     | 0.030           | 0.694          | 0.208                    | 0.080           |  |

pearance swap and generates a video without video-level artifacts such as light changes and flickers. Please refer to our supplementary videos.

#### 4.3 Ablation Studies and Further Results

Ablation studies on the expression encoding. In Tab. 3, we conduct ablation studies on different expression encoding strategies. In Direct 3DMM, we replace our CLeBS with fully-connected layers to directly inject the expression parameters of 3DMM through EAdaLN. As illustrated in Fig. 6, the direct injection does not change appearance when transferring same-identity expression however, it changes appearance (e.g., eyebrows and facial contour) when transferring cross-identity expression. Furthermore, since the raw expression parameters inherently contain noise, the generated image also exhibits visual artifacts. In **E2E LeBS**, we decrease the number of basis vectors n in LeBS for appearance bottleneck to validate the proposed contrastive pre-training. Each LeBS with n = 25, 10, 5 is jointly trained (i.e., E2E) with entire model without any pre-training. Due to the entanglement of appearance and expression, both appearance and expression are changed as a whole as the the number of basis vector n decreases. LeBS alone is insufficient for extracting appearance-free expression from the expression parameters.

Ablation studies on EAdaLN. In Tab. 4, we conduct ablation studies on EAdaLN (w. EAdaLN) by comparing it with cross-attention (w. Crossattention), which is a widely used conditioning method in transformer-based architectures [40,52]. Specifically, we replace all the EAdaLN blocks in G (Fig. 3) with cross-attention blocks. In both scenarios, CLeBS serves the refined expression  $\beta'$ . As shown in Tab. 4 and Fig. 7, the cross-attention fails to handle the expression accurately, which verifies the effectiveness of our EAdaLN for the expression conditioning.

Visualization of facial expression parameters. In Fig. 8, we sample 10 random frames from 10 different videos of distinct individual in VFHQ [59] and visualize the low-dimensional t-SNE [35] results of the two expression parameters:  $\beta \in \mathbb{R}^{64}$  and  $\beta' \in \mathbb{R}^d$ . In Fig. 8a, the 3DMM expression parameters show strong entanglement with respective to their identities, indicating the hidden appearance information in them. On the other hand, as shown in Fig. 8b, our contrastive pre-training mitigates the entanglement, thereby resolving the appearance swap in the cross-identity expression transfer in Fig. 6.



Fig. 8: Visualization of the expression parameters. We plot t-SNE [35] of raw 3DMM expression and our appearance-free expression parameter.



Fig. 9: Linear scaling along the different basis vectors of CLeBS. We visualize the different expression directions along the basis vectors  $\mathbf{v}_4, \mathbf{v}_8 \in \mathbf{V}$ .



Fig. 10: Novel-view synthesis results with expression transfer. Our method can generate more multi-view consistent images compaired to HiDe-NeRF [30].

Linear scaling along the orthogonal directions. In Fig. 9, we verify that  $\beta' \in \mathbb{R}^d$  has the orthogonal structure where the learned basis **V** determines the different expressions even if trained in unsupervised manner and the coefficients  $\{\lambda_i\}_{i=1}^n$  scale their intensities. Specifically, we visualize two orthogonal directions  $\mathbf{v}_4$  and  $\mathbf{v}_8$  and linearly scale their coefficients  $\lambda_4$  and  $\lambda_8$  from 1 to 10. As shown in Fig. 9a,  $\mathbf{v}_4$  controls eye closing and mouth closing, while Fig. 9b illustrates that  $\mathbf{v}_8$  controls mouth opening. Notably, the orthogonal basis does not influence head movements.

Novel-view synthesis with expression transfer. In Fig. 10, we compare the results of novel-view synthesis with expression transfer to those of HiDe-NeRF [30]. Both methods utilize the tri-plane and differentiable volume rendering to generate novel-view images. However, while HiDe-NeRF transfers the driving expression by deforming the generated tri-plane into a canonical tri-plane based on driving conditions [39], our method relies on the hybrid generator **G** with EAdaLN. In both same-identity and cross-identity transfer scenarios, our method synthesizes more view-consistent results, highlighting the effectiveness of our method in expression transfer without relying on deformation. Please refer to supplementary videos.

# 5 Conclusion

We presented Export3D, a 3D-aware portrait image animation model that controls the facial expression and the camera view of a source image by leveraging the driving 3DMM expression and camera parameters. Since the expression parameters are still entangled with appearance information, we proposed a contrastive pre-training framework to extract appearance-free expressions from the parameters. These refined expressions are injected into our generator through expression adaptive layer normalization (EAdaLN) that produces a tri-plane of source identity and driving expression. Finally, differentiable volume rendering renders the tri-plane into 2D images of different views. Extensive experiments show that our contrastive pre-training framework removes the appearance information from the 3DMM expression parameters, enabling our model to transfer the cross-identity expressions without undesirable appearance swap.

Limitations and future work. While our method can generate realistic face videos with driving expressions and views, it still has several limitations. First, our method cannot separately generate the foreground and background regions as the tri-plane representation construct them as a whole. Several works address this limitation by extending the tri-plane representation [4], restricting rendering points in the ray marching process [47], or leveraging the off-the-shelf segmentation model [26] to manually separate them [30, 31, 34]. Second, our method cannot control non-facial body parts (e.g., neck and shoulders) and eye gazing as they are beyond the capability of the 3DMM parameters. We plan to address these limitations for future work.

**Ethical consideration.** Since our method can generate a realistic video using a single portrait image, it has the potential for misuse, such as fake news creations. We have planned to attach visible and invisible watermarks to the generated videos and restrict the source identities for inference in research demonstration.

## References

 Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4432–4441 (2019)

- Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8296–8305 (2020)
- Amberg, B., Knothe, R., Vetter, T.: Expression invariant 3d face recognition with a morphable model. In: 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition. pp. 1–6 (2008)
- An, S., Xu, H., Shi, Y., Song, G., Ogras, U.Y., Luo, L.: Panohead: Geometry-aware 3d full-head synthesis in 360deg. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20950–20959 (2023)
- 5. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
- Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5855–5864 (2021)
- Bhattarai, A.R., Nießner, M., Sevastopolsky, A.: Triplanenet: An encoder for eg3d inversion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3055–3065 (2024)
- 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)
- 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 (CVPR). pp. 16123–16133 (2022)
- 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 (CVPR). pp. 5799–5809 (2021)
- Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML). pp. 1597–1607 (2020)
- Cheng, K., Cun, X., Zhang, Y., Xia, M., Yin, F., Zhu, M., Wang, X., Wang, J., Wang, N.: Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
- Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699 (2019)
- 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 (CVPR). pp. 10673–10683 (2022)
- 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 (CVPRW). pp. 0–0 (2019)
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

- 16 T. Ki et al.
- Egger, B., Sutherland, S., Medin, S.C., Tenenbaum, J.: Identity-expression ambiguity in 3d morphable face models. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). pp. 1–7 (2021)
- Gao, Y., Zhou, Y., Wang, J., Li, X., Ming, X., Lu, Y.: High-fidelity and freely controllable talking head video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5609–5619 (2023)
- 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)
- Gu, J., Liu, L., Wang, P., Theobalt, C.: Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In: International Conference on Learning Representations (ICLR) (2022), https://openreview.net/forum?id=iUuzzTMUw9K
- 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 Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5784–5794 (2021)
- He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9729–9738 (2020)
- Hong, F.T., Xu, D.: Implicit identity representation conditioned memory compensation network for talking head video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 23062–23072 (2023)
- Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in Neural Information Processing Systems (NeurIPS) 34, 852–863 (2021)
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8110–8119 (2020)
- Ke, Z., Sun, J., Li, K., Yan, Q., Lau, R.W.: Modnet: Real-time trimap-free portrait matting via objective decomposition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1140–1147 (2022)
- Khakhulin, T., Sklyarova, V., Lempitsky, V., Zakharov, E.: Realistic one-shot mesh-based head avatars. In: European Conference on Computer Vision (ECCV). pp. 345–362 (2022)
- Ki, T., Min, D.: Stylelipsync: Style-based personalized lip-sync video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22841–22850 (2023)
- Ko, J., Cho, K., Choi, D., Ryoo, K., Kim, S.: 3d gan inversion with pose optimization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2967–2976 (2023)
- 30. Li, W., Zhang, L., Wang, D., Zhao, B., Wang, Z., Chen, M., Zhang, B., Wang, Z., Bo, L., Li, X.: One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17969–17978 (2023)
- Li, X., De Mello, S., Liu, S., Nagano, K., Iqbal, U., Kautz, J.: Generalizable oneshot 3d neural head avatar. Advances in Neural Information Processing Systems (NeurIPS) 36 (2024)
- Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. arXiv preprint arXiv:1906.07751 (2019)

- 33. Ma, Y., Wang, S., Hu, Z., Fan, C., Lv, T., Ding, Y., Deng, Z., Yu, X.: Styletalk: One-shot talking head generation with controllable speaking styles. arXiv preprint arXiv:2301.01081 (2023)
- 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 (CVPR). pp. 16901–16910 (2023)
- Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research 9(11) (2008)
- 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)
- Min, D., Song, M., Hwang, S.J.: Styletalker: One-shot style-based audio-driven talking head video generation. arXiv preprint arXiv:2208.10922 (2022)
- Pang, Y., Zhang, Y., Quan, W., Fan, Y., Cun, X., Shan, Y., Yan, D.m.: Dpe: Disentanglement of pose and expression for general video portrait editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 427–436 (2023)
- 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 (ICCV). pp. 5865–5874 (2021)
- 40. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022)
- 41. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 484–492 (2020)
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
- 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 Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13759–13768 (2021)
- Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2287–2296 (2021)
- Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latentbased editing of real images. ACM Transactions on Graphics (TOG) 42(1), 1–13 (2022)
- Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: Generative radiance fields for 3d-aware image synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
- Shin, M., Seo, Y., Bae, J., Choi, Y.S., Kim, H., Byun, H., Uh, Y.: Ballgan: 3d-aware image synthesis with a spherical background. arXiv preprint arXiv:2301.09091 (2023)
- Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in Neural Information Processing Systems (NeurIPS) 32 (2019)

- 18 T. Ki et al.
- Sun, J., Wang, X., Wang, L., Li, X., Zhang, Y., Zhang, H., Liu, Y.: Next3d: Generative neural texture rasterization for 3d-aware head avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20991–21002 (2023)
- Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG) 40(4), 1–14 (2021)
- Trevithick, A., Chan, M., Stengel, M., Chan, E.R., Liu, C., Yu, Z., Khamis, S., Chandraker, M., Ramamoorthi, R., Nagano, K.: Real-time radiance fields for single-image portrait view synthesis. arXiv preprint arXiv:2305.02310 (2023)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS) **30** (2017)
- Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10039–10049 (2021)
- Wang, Y., Yang, D., Bremond, F., Dantcheva, A.: Latent image animator: Learning to animate images via latent space navigation. arXiv preprint arXiv:2203.09043 (2022)
- Wu, Y., Deng, Y., Yang, J., Wei, F., Chen, Q., Tong, X.: Anifacegan: Animatable 3d-aware face image generation for video avatars. Advances in Neural Information Processing Systems (NeurIPS) 35, 36188–36201 (2022)
- Xiang, J., Yang, J., Deng, Y., Tong, X.: Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2195–2205 (2023)
- 57. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems (NeurIPS) 34, 12077–12090 (2021)
- Xie, J., Ouyang, H., Piao, J., Lei, C., Chen, Q.: High-fidelity 3d gan inversion by pseudo-multi-view optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 321–331 (2023)
- Xie, L., Wang, X., Zhang, H., Dong, C., Shan, Y.: Vfhq: A high-quality dataset and benchmark for video face super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 657–666 (2022)
- Yin, F., Zhang, Y., Cun, X., Cao, M., Fan, Y., Wang, X., Bai, Q., Wu, B., Wang, J., Yang, Y.: Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In: European Conference on Computer Vision (ECCV). pp. 85–101 (2022)
- Yin, F., Zhang, Y., Wang, X., Wang, T., Li, X., Gong, Y., Fan, Y., Cun, X., Shan, Y., Oztireli, C., Yang, Y.: 3d gan inversion with facial symmetry prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 342–351 (2023)
- Yu, W., Fan, Y., Zhang, Y., Wang, X., Yin, F., Bai, Y., Cao, Y.P., Shan, Y., Wu, Y., Sun, Z., et al.: Nofa: Nerf-based one-shot facial avatar reconstruction. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–12 (2023)
- Yuan, Z., Zhu, Y., Li, Y., Liu, H., Yuan, C.: Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. arXiv preprint arXiv:2303.12326 (2023)

- 64. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8652–8661 (2023)
- Zhao, J., Zhang, H.: Thin-plate spline motion model for image animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3657–3666 (2022)
- Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2223–2232 (2017)