

Everything's Talkin': Pareidolia Face Reenactment

Linsen Song^{1,2*} Wayne Wu^{3,4*} Chaoyou Fu^{1,2} Chen Qian³ Chen Change Loy⁴ Ran He^{1,2†}

¹School of Artificial Intelligence, University of Chinese Academy of Sciences

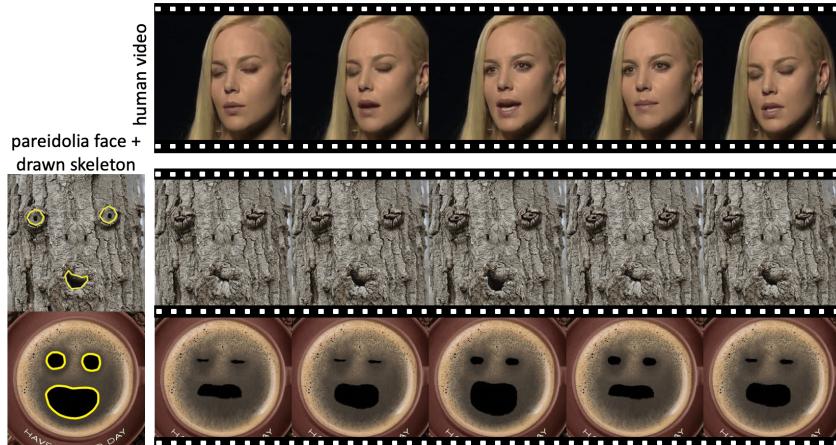
²NLPR & CRIPAC, CASIA ³SenseTime Research ⁴S-Lab, Nanyang Technological University

songlinsen2018@ia.ac.cn, {wuwenyan, qianchen}@sensetime.com,

{chaoyou.fu, rhe}@nlpr.ia.ac.cn, ccloy@ntu.edu.sg



(a) Examples of pareidolia faces



(b) Pareidolia face reenactment

Figure 1: **Pareidolia Face Reenactment.** We propose an unsupervised method for static illusory faces to become animated by reenacting the human face in a video. The eyes and mouth of the illusory faces, as defined by users, move in tandem with those of the human in the video, simultaneously.

Abstract

We present a new application direction named *Pareidolia Face Reenactment*, which is defined as animating a static illusory face to move in tandem with a human face in the video. For the large differences between pareidolia face reenactment and traditional human face reenactment, two main challenges are introduced, i.e., shape variance and texture variance. In this work, we propose a novel *Parametric Unsupervised Reenactment Algorithm* to tackle these two challenges. Specifically, we propose to decompose the reenactment into three catenate processes: shape modeling, motion transfer and texture synthesis. With the decomposition, we introduce three crucial components, i.e., *Parametric Shape Modeling*, *Expansionary Motion Transfer* and *Unsupervised Texture Synthesizer*, to overcome the problems brought by the remarkably variances on pareidolia faces. Extensive experiments show the superior performance of our method both qualitatively and quantitatively. Code, model and data are available on our project page¹.

*Equal contribution.

†Corresponding author.

¹<https://wywu.github.io/projects/ETT/ETT.html>

1. Introduction

It's not often that you look at your meal to find it staring back at you. But when Diane Duyser picked up her cheese toastie, she was in for a shock. "I went to take a bite out of it, and then I saw this lady looking back at me," she told the Chicago Tribune. "It scared me at first." [1].

The phenomenon described in this BBC news is called face pareidolia, a natural inclination of the human brain to perceive illusory faces that do not actually exist [19, 50]. In this work, we attempt to bring this interesting imagination into reality by animating. As shown in Fig. 1 (b), we propose a new application direction named "*Pareidolia Face*" *Reenactment*, which is defined as animating illusory faces by the motion extracted from human faces automatically.

Pareidolia face reenactment, has large potential usages in filmmaking [46, 24], cartoon production [53, 56] and mixed reality [47, 55], which always requires a massive labor of professional animators. Mostly related, face reenactment [13, 46, 25, 52] is becoming an emerging topic in recent years. However, all of these methods are designed specifically for human faces, of which rich priors like facial

landmarks [54, 18] or 3D face models [46, 24] can be utilized. But, all of these priors are unachievable for pareidolia faces. Moreover, large-scale face datasets [5, 35] with massive annotations are sufficient for human faces, which are also unreachable for pareidolia faces. Reenacting pareidolia faces by human portrait videos is still an open question.

The main challenges for pareidolia face reenactment can be summarized into two large *variances*, *i.e.*, shape variance and texture variance. *Shape variance* means that the boundary shapes of facial parts are remarkably diverse, such as circular, square and moon-shape mouths as shown in Fig. 1 (a). For human faces, landmarks are always used as the intermediary to transfer motions [7, 18, 56]. However, landmark suffers from the tightly coupling with the shape/size of facial parts. It cannot be used as the intermediary to perform a precise motion transfer from the source human face to the target pareidolia face. The shapes of target faces will be affected by the source ones’ easily. Also, it is difficult to define the meaning of landmarks’ annotation for complex shapes, *e.g.*, the tree’s mouth in Fig. 1 (b). Thus, it is challenging to design a universal shape representation to transfer motion from human faces to pareidolia faces.

Texture variance means the textures of pareidolia faces are remarkably diverse, such as wood, downy and metal textures as shown in Fig. 1 (a). Also, the texture distribution of pareidolia faces is extremely discrete, since there is even no two faces with a similar texture. For human face, previous works always deployed 3D facial models [46, 24] or GAN-based generator [18, 54] in texture synthesis. However, there is no 3D face model that can be leveraged to model pareidolia face. Also, for the GAN-based synthesis, large-scale labeled datasets with landmark-image pairs are always needed to train a generator [32, 52]. But, there exists no dataset or annotation for pareidolia faces, which makes the strong supervision with paired data out of action for texture synthesis. Thus, synthesizing the textures of pareidolia faces is challenging, without a 3D model or annotated data.

In this work, we propose a novel Parametric Unsupervised Reenactment Algorithm, to tackle the pareidolia face reenactment problem. First, to solve the shape variance challenge, we propose a *Parametric Shape Modeling* technique, in which we introduce Bézier Curve [9], a classic parametric technology in computer graphics, to represent the boundary shapes of facial parts of both the source and target faces with a set of control points. With the parametric modeling of boundaries on target pareidolia face, control points of the Bézier Curve can locally modify the curve while keeping its global structure unchanged, even with large shape variance.

With the robust shape representation, a naïve solution to transfer motion is to directly adapt the human face’ control points to the pareidolia face. However, the transferred motion so far only decides the movement of the facial bound-

aries of the pareidolia face, which is a local motion and cannot be used to drive the whole face. Thus, we propose an *Expansionary Motion Transfer* technique to get a global motion representation named motion field for a natural animation, in which a Motion Spread strategy is designed to propagate the transferred motion from boundary to the whole face and a First-order Motion Approximation strategy is designed to refine the motion field further.

While the motion has been successfully transferred, the next step is to use the motion field to deduce an image with high-quality textures. Reviewing the challenge of texture variance, we propose an *Unsupervised Texture Synthesizer* to address it in an AutoEncoder framework with a carefully designed Feature Deforming Layer. High-quality textures can be synthesized successfully, while neither 3D model, nor large-scale face datasets with annotations are needed.

We summarize our contributions as follows: 1) We make the first attempt to animate pareidolia faces by the facial motion derived from the human faces. 2) We propose a novel Parametric Unsupervised Reenactment Algorithm to tackle pareidolia face reenactment, with three crucial components, *i.e.*, Parametric Shape Modeling, Expansionary Motion Transfer and Unsupervised Texture Synthesizer. 3) Extensive experiments present the superior performance of our method and the effectiveness of each component.

2. Related Work

2.1. Face Reenactment

Face reenactment refers to transferring motion patterns from one face to another one, including both graphics-based [45, 2] and learning-based [18, 22, 32, 43] methods. The former mainly relies on 3DMMs [4]. Benefiting from the face fitting capacity of 3DMMs, recent methods, *e.g.* Face2Face [46] and DVP [25], can reenact a given face by adjusting the fitted parameters. Nevertheless, 3DMMs, designed for human faces, are inapplicable to our pareidolia faces. The latter mainly resorts to Deep Neural Networks (DNNs). Thanks to the powerful expression capacity of DNNs, GAN [16]-based methods [21, 40, 10, 12, 11, 49] like ReenactGAN [52] can achieve face reenactment via learning a mapping from a source face to a target one. However, these methods usually require a large amount of paired training data, which are unavailable in our task. Furthermore, apart from reenacting human faces, there are also some methods try to animate non-human faces, such as the cat face [51, 33] or the cartoon face [56]. Recent audio-driving method [56] animates cartoon faces while their labeled 68 facial landmarks correspond to human faces. These methods are all driven by facial landmarks, which are inapplicable in our pareidolia face reenactment task.

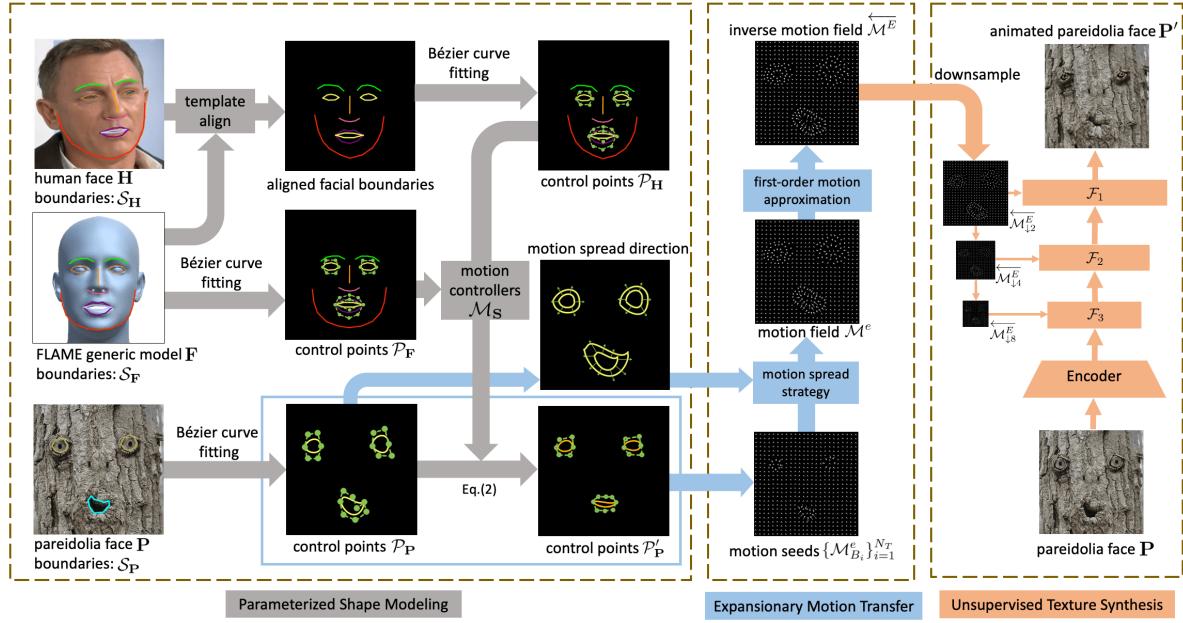


Figure 2: **Parametric Unsupervised Reenactment Algorithm.** We separate the proposed algorithm into three components: (a) Parametric Shape Modeling. First, we model the facial boundary of three kinds of faces on the left-most column, with Bézier Curve fitting. Then the facial boundaries can be represented as a set of control points. By simply adapting the motion controller inferred from the human and FLAME face, we can get the animated control points of the pareidolia face \mathbf{P}' . (b) Expansionary Motion Transfer: The animated control points are then converted into an optical flow map as motion seeds, which represent a local motion of the target pareidolia face. Motion Spread and First-order Motion Approximation are proposed to extend the local motion seed to a global motion field \mathcal{M}^e . (c) Unsupervised Texture Synthesizer: With the motion field and the raw pareidolia face as conditions, we can synthesize the final animated pareidolia face.

2.2. Geometric Shape Modeling

There are two means to model facial geometric shape: implicit or explicit modeling. The former directly disentangles shape representations from faces via elaborately designed networks and training manners [48, 27, 6]. The latter leverages extra auxiliary models to present facial shape information, e.g. facial landmarks [21, 54, 14] or 3D parameters [44, 41]. Bézier curves are a very popular tool in computer-aided design [15], computer graphics and interactive curve design [36]. Recently, Bézier curves are incorporated with deep learning methods like CNNs [26] and GANs [16] in tasks includes parametric skeleton extraction [30] and sketch generation from human drawing [42]. In this paper, we explore the application of Bézier curves in modeling the boundary of facial parts.

3. Methodology

The architecture of our proposed method is shown in Fig. 2, which is separated into three main components: Parametric Shape Modeling, Expansionary Motion Transfer and Unsupervised Texture Synthesizer. First, we extract the boundaries of both human and pareidolia faces. We build a robust shape model for the facial boundaries based on the *Bézier Curve* and represent the motion as the *Motion Con-*

troller (Sec. 3.1). Then, we get an optical flow map named motion seed to represent the transferred local motion at the facial boundaries. In order to animate the whole pareidolia face, we propose *Motion Spread* and *First-order Motion Approximation* strategy to induce a global motion representation named motion field. (Sec. 3.2). At last, we propose an unsupervised network with a carefully designed *Feature Deforming Layer* to synthesize high-quality animated texture conditioned on the static pareidolia face and the motion field. (Sec. 3.3)

3.1. Parametric Shape Modeling

Reviewing the remarkable shape variance exists in facial parts’ shapes of human and pareidolia faces. Bézier Curve can be edited locally while remain the whole structure. Thus, we introduce it to robustly model the shapes of facial parts. In this section, we first describe the Shape Modeling of facial boundaries in detail, then introduce Motion Controller, the motion representation based on our shape modeling fashion.

Shape Modeling by Bézier Curves. Composite Bézier Curve [38], defined as a piecewise Bézier Curve [9], is exploited to fit the facial boundaries since composite Bézier Curves can freely model complex boundary and each of its control points can regulate the curve locally and do not

break the curve’s global structure.

As shown in Fig. 2, to represent the human face’s shape independent of the face scale and rotation, we use the template alignment algorithm [37] to affine the human face \mathbf{H} to the referred generic head model \mathbf{F} of FLAME [28]. The aligned facial boundaries will be used for the following shape modeling. In Fig. 3, we illustrate the procedure of shape modeling for human facial parts in detail. First, by connecting inferred 68 3D landmarks [17] of the human face \mathbf{H} we obtain facial parts’ boundaries $\mathcal{S}_{\mathbf{H}} = \{C_i^{\mathbf{H}}\}_{i=1}^N$ composed of N_H branches. For example, the mouth boundary can be divided into four branches: inner and outer contours of upper and lower lips. Then, each branch $C_i^{\mathbf{H}}$ is fitted by a single h_i -order composite Bézier Curve $B_i^{\mathbf{H}}$ parameterized by $h_i + 1$ control points $\mathcal{P}_{\mathbf{H}}^{B_i} = \{(\hat{x}_j^i, \hat{y}_j^i, \hat{z}_j^i)\}_{j=0}^{h_i}$, where $(\hat{x}_j^i, \hat{y}_j^i, \hat{z}_j^i)$ represents 3D coordinates of each control point. Please refer to the *supplementary material* about the single n -order composite Bézier curve fitting. Thus, all branches of $\mathcal{S}_{\mathbf{H}}$ can be parameterized as control points $\mathcal{P}_{\mathbf{H}} = \bigcup_{i=1}^N \mathcal{P}_{\mathbf{H}}^{B_i}$. At last, we conduct a similar procedure on the referred FLAME head model \mathbf{F} and the boundaries $\mathcal{S}_{\mathbf{F}} = \{C_i^{\mathbf{F}}\}_{i=1}^N$ are parameterized as control points $\mathcal{P}_{\mathbf{F}} = \bigcup_{i=1}^N \mathcal{P}_{\mathbf{F}}^{B_i} = \bigcup_{i=1}^N \{(\bar{x}_j^i, \bar{y}_j^i, \bar{z}_j^i)\}_{j=0}^{h_i}$. For a pareidolia face, we manually label its boundaries $\mathcal{S}_{\mathbf{P}} = \{C_i^{\mathbf{P}}\}_{i=1}^{N_P}$ composed of N_P branches and they are parameterized of control points $\mathcal{P}_{\mathbf{P}}$. By now, we get control points of facial parts’ boundaries for \mathbf{H} , \mathbf{F} and \mathbf{P} as $\mathcal{P}_{\mathbf{H}}$, $\mathcal{P}_{\mathbf{F}}$, $\mathcal{P}_{\mathbf{P}}$ respectively, which is used for the following motion controller’s calculation.

Motion Controller. Our motion representation extracted from human face, denoted as $\mathcal{M}_{\mathbf{S}}$, is defined as position of control points $\mathcal{P}_{\mathbf{H}}$ relative to $\mathcal{P}_{\mathbf{F}}$ as follows:

$$\mathcal{M}_{\mathbf{S}} = \bigcup_{i=1}^N \left\{ \left(\frac{\hat{x}_j^i}{\bar{x}_j^i}, \frac{\hat{y}_j^i}{\bar{y}_j^i}, \frac{\hat{z}_j^i}{\bar{z}_j^i} \right) \right\}_{j=0}^{h_i}, \quad (1)$$

$\mathcal{M}_{\mathbf{S}}$ will animate boundaries of pareidolia faces $\mathcal{S}_{\mathbf{P}}$ and it is called as *motion controllers* in Fig. 3. In general, boundary branches of a pareidolia face are a subset of those of a human face, *e.g.*, $N_P \leq N$. Because some facial parts of the pareidolia face such as nose and jawline are hard to define and only the observed facial parts are adopted, *e.g.*, eyes and mouth in Fig. 2. First, for simplicity, we assume that the i -th boundary branch of \mathbf{P} is corresponding to the i -th boundary branch of \mathbf{H} . We parametrize the boundaries $\mathcal{S}_{\mathbf{P}}$ by control points $\mathcal{P}_{\mathbf{P}} = \bigcup_{i=1}^{N_P} \mathcal{P}_{B_i} = \bigcup_{i=1}^{N_P} \{(\hat{x}_j^i, \hat{y}_j^i, \hat{z}_j^i)\}_{j=0}^{t_i}$, where N_P is the branch number and t_i is the curve order of the i -th branch. Then, we note that the shape of curve B_i (pareidolia face) might greatly differ from the shape of corresponding curve $B_i^{\mathbf{H}}$ (human face), *e.g.*, the number of control points differs ($t_i \neq h_i$). Thus, we uniformly remove (when $t_i < h_i$) or linearly interpolate (when $t_i > h_i$) the ordered motion controllers of the curve $B_i^{\mathbf{H}}$ in $\mathcal{M}_{\mathbf{S}}$ and de-

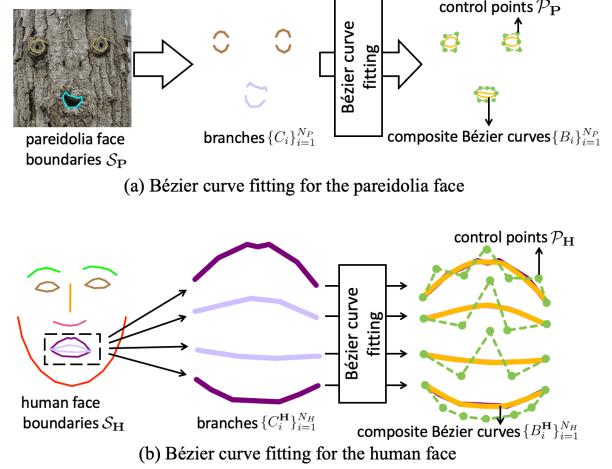


Figure 3: **Parametric Shape Modeling.** The boundaries of human face ($\mathcal{S}_{\mathbf{H}}$) and pareidolia face ($\mathcal{S}_{\mathbf{P}}$) are parameterized as control points of composite Bézier curves. For simplicity, we use the mouth as an example in $\mathcal{S}_{\mathbf{H}}$.

note the adapted motion controllers as $\mathcal{M}_{\mathbf{S}}^e$. At last, facial parts of pareidolia face are animated by applying the motion controllers $\mathcal{M}_{\mathbf{S}}^e$ on their control points $\mathcal{P}_{\mathbf{P}}$ as follows:

$$\mathcal{P}'_{\mathbf{P}} = \mathcal{M}_{\mathbf{S}}^e \otimes \mathcal{P}_{\mathbf{P}} = \bigcup_{i=1}^{N_P} \left\{ \left(\frac{\hat{x}_j^i}{\bar{x}_j^i} x_j^i, \frac{\hat{y}_j^i}{\bar{y}_j^i} y_j^i, \frac{\hat{z}_j^i}{\bar{z}_j^i} z_j^i \right) \right\}_{j=0}^{t_i}, \quad (2)$$

where \otimes is point-wise dot product and $\mathcal{P}'_{\mathbf{P}}$ is the control points of the boundaries animated by motion $\mathcal{M}_{\mathbf{S}}^e$.

3.2. Expansionary Motion Transfer

Now we transfer the motion at the facial boundaries by animated control points of composite Bézier curves in the pareidolia face. However, the transferred motion is local and a global motion of the whole face is required to animate a pareidolia face. Thus, we develop a Motion Spread strategy to expand the motion at the boundaries to the whole face. Moreover, we find that the texture animated by the expanded motion contains missing pixels as shown in Fig. 5 (a). We will detail the cause and then propose the First-order Motion Approximation to address it.

Motion Spread. The motion at curve B_i , denoted as $\mathcal{M}_{B_i}^e$, is defined by the optical flow map of points on composite Bézier Curves parameterized by $\mathcal{P}_{\mathbf{P}}$ and $\mathcal{P}'_{\mathbf{P}}$. We call the facial motion at each boundary branch of facial parts, *e.g.* $\mathcal{M}_{B_i}^e$, as the *motion seed*. Then $\{\mathcal{M}_{B_i}^e\}_{i=1}^{N_P}$ is called as *motion seeds* as shown in Fig. 2. Each composite Bézier curve B_i is related to a motion seed $\mathcal{M}_{B_i}^e$ in the pareidolia face \mathbf{P} . Note that the motion seeds only define a very local motion at the facial boundaries, we develop a *Motion Spread* strategy to derive the facial motion of the whole pareidolia face.

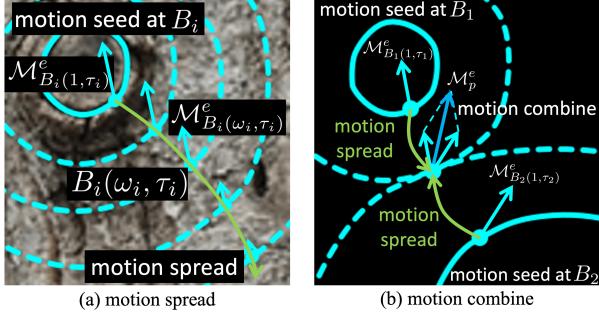


Figure 4: (a) **motion spread**. A motion $\mathcal{M}_{B_i(1, \tau_i)}^e$ of the motion seed at B_i spreads to $B_i(\omega_i, \tau_i)$ as motion $\mathcal{M}_{B_i(\omega_i, \tau_i)}^e$. (b) **motion combine**. At pixel p , we combine spreaded motion from motion seeds at B_1, B_2 .

Our Motion Spread strategy decays motion seeds along the directions orthogonal with their composite Bézier curves as shown in Fig. 4 (a). Each motion seed $\mathcal{M}_{B_i}^e$ represents the motion of all points on curve B_i . We expand the curve B_i to different scales to cover its neighboring area. If a pixel position p locates at the relative position τ_i ($\tau_i \in [0, 1]$) of ω_i -time scaled curve B_i , we can represent it as $p = B_i(\omega_i, \tau_i)$. Then, the decayed motion $\mathcal{M}_{B_i(\omega_i, \tau_i)}^e$ from the motion seed $\mathcal{M}_{B_i}^e$ can be written as follows:

$$\mathcal{M}_{B_i(\omega_i, \tau_i)}^e = \lambda(\omega_i) \cdot \mathcal{M}_{B_i(1, \tau_i)}^e, \quad \mathcal{M}_{B_i(1, \tau_i)}^e \in \mathcal{M}_{B_i}^e, \quad (3)$$

where \cdot means scalar multiplication and the motion decay factor $\lambda(\omega_i)$ is determined by ω_i as presented in the *supplementary material*. One pixel p in the pareidolia face might receive decayed motion from several motion seeds. Therefore, we calculate the motion at p by motion combine. The motion at p , denoted as \mathcal{M}_p^e , is the combination of these decay motion as shown in Fig. 4 (b).

To animate the pareidolia face \mathbf{P} , a global motion for the whole pareidolia face is built from the motion seeds through our proposed Motion Spread strategy. Such a global motion is constituted by the motion of \mathcal{M}_p^e for all $p \in G(\mathbf{P})$, where $G(\cdot)$ is a function that returns the pixel grid of the input image. We call it as *motion field* and denote it as $\mathcal{M}^e = \{\mathcal{M}_p^e\}_{p \in G(\mathbf{P})}$. Then, for a pixel at p in the pareidolia face, the motion field can animate it to a new location $\mathcal{M}_p^e + \mathbf{p}$ ($\forall p \in G(\mathbf{P}), \mathcal{M}_p^e \in \mathcal{M}^e$) by the motion field \mathcal{M}^e .

First-order Motion Approximation. The motion field \mathcal{M}^e can be used to animate the pareidolia face \mathbf{P} as the reenacted face \mathbf{P}' . If we regard \mathcal{M}^e as a function $G(\mathbf{P}) \rightarrow G(\mathbf{P}')$, then it is neither an injection nor a surjection since multiple pixel locations in $G(\mathbf{P})$ might be mapped to one pixel location in $G(\mathbf{P}')$. Thus, directly using \mathcal{M}^e to animate \mathbf{P} will cause some missing pixels in \mathbf{P}' as shown in Fig. 5 (a). To solve this problem, we introduce the inverse function of \mathcal{M}^e , denoted as $\overleftarrow{\mathcal{M}}^e : G(\mathbf{P}') \rightarrow G(\mathbf{P})$ since $\overleftarrow{\mathcal{M}}^e$ do not have valid function value at the locations

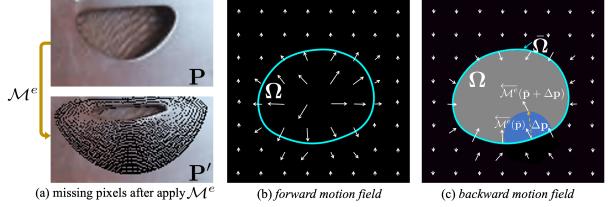


Figure 5: **First-order Motion Approximation.** (a) A practical example of missing pixels after directly applying the motion field. (b) *Motion field* represents the motion of $G(\mathbf{P}) \rightarrow G(\mathbf{P}')$. The pixel inside the hole Ω move outward. (c) *Inverse motion field* misses valid values inside Ω and the proposed First-order Motion Approximation is used to approximate them.

of missing pixels. Thus, our goal is to inpaint the $\overleftarrow{\mathcal{M}}^e$ and we propose our *First-order motion approximation*.

To illustrate the First-order Motion Approximation, we take a typical case in Fig. 5 (b) where all pixels in the area Ω move outward. At first, we call $\overleftarrow{\mathcal{M}}^e$ as *inverse motion field* since it defines the pixel movements opposite to that of the motion field. Then, in Fig. 5 (c), $\overleftarrow{\mathcal{M}}^e$ does not contain valid function value in the area $\Omega \subset G(\mathbf{P}')$. At last, since close pixels will have similar motion, we propose the First-order Motion Approximation for $\overleftarrow{\mathcal{M}}^e$ to spread facial motion from the area boundary $\bar{\Omega}$ to the area Ω . Specially, We compute the first-order Taylor expansion of $\overleftarrow{\mathcal{M}}^e$ around the area boundary pixel $\bar{p} \in \bar{\Omega}$ as follows:

$$\overleftarrow{\mathcal{M}}^e(\bar{p} + \Delta p) = \overleftarrow{\mathcal{M}}^e(\bar{p}) + \left(\frac{d}{dp} \overleftarrow{\mathcal{M}}^e(p) \Big|_{p=\bar{p}} \right) \Delta p + o(||\Delta p||), \quad (4)$$

where $\bar{p} + \Delta p$ is a nearby pixel of the boundary \bar{p} . We omit $o(||\Delta p||)$ of Eq. (4) to approximate $\overleftarrow{\mathcal{M}}^e(\bar{p} + \Delta p)$. We conduct Eq. (4) for all missing pixels in $\overleftarrow{\mathcal{M}}^e$ and the enhanced *inverse motion field* is denoted as $\overleftarrow{\mathcal{M}}^E$. The motion approximation along Δp can be viewed in Fig. 5 (c).

3.3. Unsupervised Texture Animator

While we get the inverse motion field, the final step is to synthesize the image result conditioned on the raw pareidolia face \mathbf{P} and the inverse motion field. Reviewing the absence of large-scale datasets and annotations for pareidolia face, we propose an AutoEncoder based Unsupervised Texture Animator. Specifically, we first train a simple AutoEncoder with only natural images without any annotation. The trained AutoEncoder can be seen as a texture reconstructor by now, which can reconstruct the input texture but cannot animate it. To this end, we design a Feature Deforming Layer, which is coupled with the AutoEncoder network, to transfer the motion to texture progressively. Note that Feature Deforming Layer is only used in the inference stage, which makes the training unsupervised and enjoys the diversity of large-scale datasets of natural images.

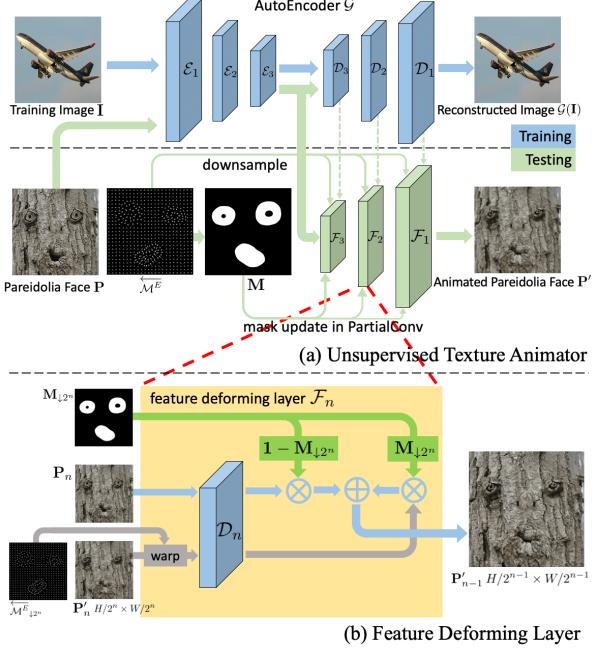


Figure 6: (a) **Unsupervised Texture Animator.** Training and testing phases of our Unsupervised Texture Animator. (b) **Feature Deforming Layer.** For simplifying the understanding, we visualize the feature maps P_n, P'_n, P'_{n-1} using the tree texture images.

Unsupervised AutoEncoder. We train an unsupervised AutoEncoder \mathcal{G} to extract image features at different scales as shown in Fig. 6. During the training phase, an image I is fed into \mathcal{G} to produce reconstructed image $\mathcal{G}(I)$. We apply l_1 reconstruction loss L_{rec} and perceptual loss [23] L_{vgg} on I and $\mathcal{G}(I)$. The loss function of \mathcal{G} is $L_{\mathcal{G}} = \alpha_1 L_{rec} + \alpha_2 L_{vgg}$, where α_1 and α_2 are set empirically. We put the network details in the *supplementary material*.

Feature Deforming Layer. The pareidolia image features P_n of the scale $H/2^n \times W/2^n$ can be retrieved from layer D_n in the pretrained \mathcal{G} . During the testing phase, in Fig. 6, we design the *Feature Deforming Layer* \mathcal{F}_n to warp the synthesized features P'_n by the downsampled motion field $\overleftarrow{\mathcal{M}}^E_{\downarrow 2^n}$ and refine it by D_n . Thus, the texture is progressively synthesized by $\mathcal{F}_3, \mathcal{F}_2, \mathcal{F}_1$. At last, since some pixels of P do not move in P' . A 0-1 motion mask M is calculated based on $\overleftarrow{\mathcal{M}}^E$ to keep the texture at these pixels unchanged. The motion mask M is defined as $M = \mathbb{1}(\overleftarrow{\mathcal{M}}^E - \mathcal{M}^I)$, where $\mathbb{1}$ is an indicator function that returns 0 if the input is 0 and returns 1 elsewhere, and \mathcal{M}^I is an identity motion field that do not move any pixel of P . Thus, the texture refinement of layer \mathcal{F}_n is written as:

$$P'_{n-1} = \mathcal{F}_n(P_n, P'_n, \overleftarrow{\mathcal{M}}^E_{\downarrow 2^n}, M_{\downarrow 2^n}), \quad (5)$$

where $M_{\downarrow 2^n}$ is obtained by downsampling the mask M (2^n is the scale factor) through the mask update method in PartialConv [31]. In Eq. (5), P'_3 is our coarsest texture produced by the encoder of \mathcal{G} . By the progressive warp and

refine, we get $P'_0 = P'$ as our final synthesized pareidolia face with the same motion of the human face video.

4. Experiments

We show qualitative and quantitative results on the generated videos of pareidolia faces to demonstrate the performance of our reenactment method for pareidolia faces and the effectiveness of the proposed components.

Datasets. During the training phase, the AutoEncoder \mathcal{G} is trained on the COCO2017 dataset [29]. During the testing phase, the human portrait videos include videos from *Obama Weekly Address* [43] and CelebVox2 [8]. Moreover, we collect a dataset PareFace, which includes 1,000 pareidolia faces to facilitate future researches on this topic. More samples can be viewed in our *supplementary material*.

Metric. We evaluate videos of the animated pareidolia faces in terms of textures and the shape/motion of facial parts. We use IS [3], FID [20] to evaluate the synthesized texture quality. Due to the lack of metrics about evaluating the shape and motion differences between human and pareidolia faces. We design the following metrics to evaluate the shape similar and motion accuracy: 1) **Shape similarity (S-Sim).** Following [34], we use the eccentricity histogram of a shape as its descriptor. The cosine distance is used to measure the similarity between two shapes. 2) **Close-open accuracy (CO-Acc).** To measure the extreme motion accuracy of the mouth and eyes, we compare their open/close status. The CO-Acc is defined as the average difference of the mouth/eyes open ratio between input human and animated pareidolia faces, where the open ratio is expressed as a percentage of the maximum height of the mouth/eyes. 3) **Motion accuracy (M-Acc).** To measure the overall motion accuracy of the mouth and eyes, we compare their tendencies of becoming larger/smaller. We use 0-1 flags to denote if the area becomes larger or smaller in the next frame. The flag serves as a motion indicator and we compare its average differences between human and pareidolia faces.

4.1. Pareidolia Face Reenactment

We use a human portrait video to drive a given pareidolia face and present visual results in Fig. 7. It can be seen that the generated pareidolia face imitates the motion of the input human face at the mouth and eyes areas, even the subtle size changing. Benefits from our parametric shape modeling, the prominent motion at facial parts is transferred from the human to pareidolia face. Our Motion Spread strategy makes it possible to animate the area around facial parts and make the whole pareidolia face looks more lively. Even large texture discrepancy exists between human and pareidolia faces as shown in Fig. 8 (a), where the distribution of pareidolia faces' texture is more discrete than that of human faces. We recommend to view reenactment results in the *supplementary video*.

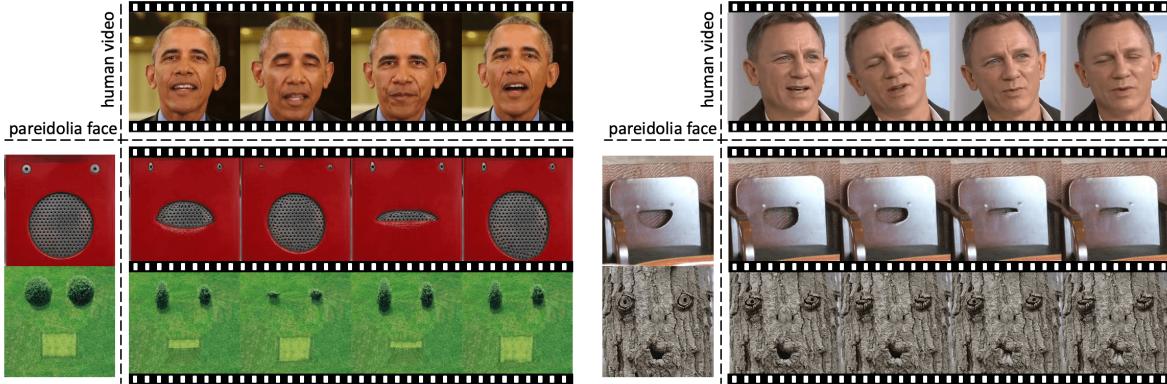


Figure 7: **Pareidolia Face Reenactment.** In each block, we use the human face portrait video in the first row to drive pareidolia faces in the second and third row. Our method mainly focuses on facial motion transferring at the mouth and eyes.

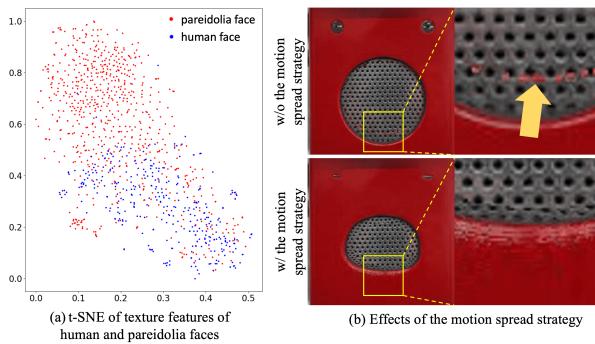


Figure 8: (a) **t-SNE of texture features of human and pareidolia faces.** The texture feature is extracted by VGG16 [39] pre-trained on ImageNet. (b) **Effect of Motion Spread Strategy.** Comparison of animated results before and after applying the Motion Spread strategy.

4.2. Ablation Study

In this section, we present an ablation to evaluate the effectiveness of our proposed components. First, we manually label “landmarks” for pareidolia faces and use them to replace the control points of composite Bézier curves to validate the effect of our parametric shape modeling (Sec. 4.2.1). Then, we show the decisive role that the Motion Spread strategy plays in the pareidolia face animation (Sec. 4.2.2). In addition, we compare the texture quality improvement brought by our First-order Motion Approximation (Sec. 4.2.3). At last, we progressively add our proposed Feature Deforming Layers ($\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$) in the Unsupervised Texture Animator to visualize the progressively refined textures (Sec. 4.2.4).

4.2.1 Composite Bézier Curve v.s. Landmarks

To valid the superiority of incorporating composite Bézier curves, we compare our method with the one that replaces the control points of composite Bézier curve with the manually labeled “landmarks” for pareidolia faces. We present

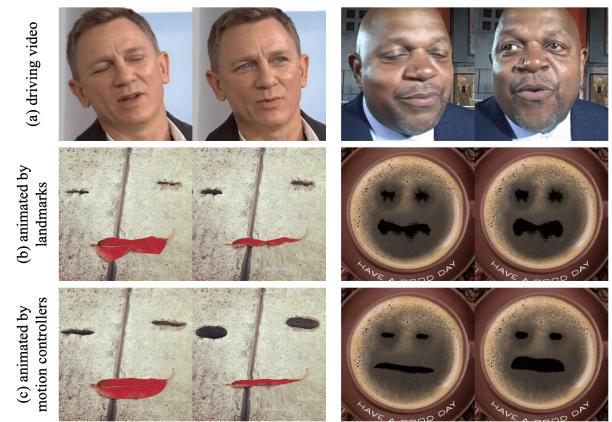


Figure 9: **Composite Bézier Curves v.s. Landmarks.** Our proposed representation *motion controllers* inspired by composite Bézier curve is better than landmarks in reenacting pareidolia faces.

the qualitative results in Fig. 9. We can see that the facial parts’ global shape is broken when landmarks are applied while our method preserves them well. Also, the reenactment results produced by our method also imitate the facial motion of the human face better (the eyes’ motion of the left man in Fig. 9). To quantitatively compare the shape similar, motion accuracy and image quality, we present Tab. 1 (a) and Tab. 1 (b) that compare the results of driving by landmarks and composite Bézier curves. Compare with landmarks, modeling facial parts by composite Bézier curves are better at preserving the facial parts’ global shapes, imitating the motion of human faces and the synthesized image visual quality.

4.2.2 Effect of the Motion Spread Strategy

We propose the Motion Spread strategy to obtain the motion field that defines the global motion of the pareidolia face from the motion seeds that only define the motion of facial parts’ boundaries. As shown in Fig. 8 (b), only the pixels at

Table 1: (a) **Shape similarity and motion accuracy of Motion Controllers v.s. Landmarks.** In the table, 'm' means mouth and 'e' means eyes. (b) **Visual Quality Comparison.** The IS and FID of images synthesized by different settings are compared. In the 'landmark', we use labeled landmarks of pareidolia faces instead of composite Bézier curves. In the 'w/o motion appr.', we do not apply the First-order Motion Approximation. In the 'Ours', the Unsupervised Texture Animator applies the network structure $\mathcal{F}_2, \mathcal{F}_2, \mathcal{F}_3$.

(a) Method	landmark	composite Bézier curve
S-Sim(m)	0.43	0.75
S-Sim(e)	0.55	0.82
CO-Acc(m)	0.52	0.76
CO-Acc(e)	0.71	0.82
M-Acc(m)	0.77	0.84
M-Acc(e)	0.80	0.89

(b) Method	IS	FID
landmark	8.21	13.1
$\mathcal{F}_1, \mathcal{D}_2, \mathcal{D}_3$	8.79	12.9
$\mathcal{F}_2, \mathcal{F}_2, \mathcal{D}_3$	8.89	12.5
w/o motion appr.	9.17	12.3
Ours	9.22	12.3

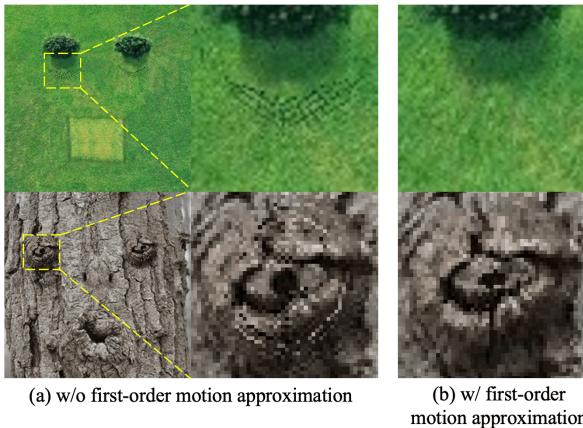


Figure 10: **Effect of First-order Motion Approximation.** The texture synthesized by our Unsupervised Texture Animator when our proposed First-order Motion Approximation is applied (b) or not (a).

facial boundaries are animated without the Motion Spread strategy, which presents failed animation results. Benefiting from the Motion Spread strategy, we can animate the whole pareidolia face.

4.2.3 Effect of the Motion Approximation

Our First-order Motion Approximation is designed to approximate the missing motion in the inverse motion field. Thus, the quality of the synthesized textures is improved after introducing this motion approximation method, as shown in Fig. 10. Some pixel-level visual artifacts in Fig. 10 (a) are caused by the missing value in the inverse motion field, thus the original pixels are remained. They are obviously eliminated after applying our First-order Motion Approximation. We also present the quantitative results of image quality in Tab. 1 (b). Both IS and FID are slightly improved after applying our motion approximation method.

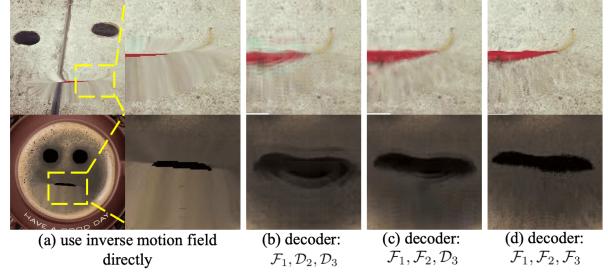


Figure 11: **Effect of the Feature Deforming Layers.** (a) We directly use the inverse motion field to animate the pareidolia face. (b-d) The texture quality of animated pareidolia faces is progressively improved by our Feature Deforming Layers.

4.2.4 Effect of the Feature Deforming Layer

Directly applying the inverse motion field usually induce blurred synthesized textures. Our Feature Deforming Layers in the Unsupervised Texture Animator benefits the texture synthesis by progressively warp and refine the input pareidolia faces. We present the textures synthesized by each Feature Deforming Layer in Fig. 11 where we can see that the synthesized textures become more and more clear after we add $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ layer by layer. We also present the quantitative results of image quality in terms of IS and FID in Tab. 1 (b) where both metrics are improved if more Feature Deforming Layers are applied.

5. Discussion

In this paper, we make the first attempt on pareidolia face reenactment, which might benefit the cartoon production and mixed reality in the future. We present a Parametric Unsupervised Reenactment Algorithm to tackle this challenging problem and demonstrate superior reenactment results in the experiments.

Pareidolia face reenactment is an extremely challenging problem that might fail when meet extreme head poses, complex shapes and textures of the pareidolia faces. Our method mainly focuses on transferring the motion of eyes and mouth from human faces to the frontal pareidolia faces. We leave the motion transferring of other facial parts (*e.g.* nose and facial muscles), the head movement and animating non-frontal pareidolia faces as our future works.

Acknowledgments

This research was conducted in collaboration with SenseTime. This work is partially funded by Beijing Natural Science Foundation (Grant No. JQ18017), Youth Innovation Promotion Association CAS (Grant No. Y201929), and National Natural Science Foundation of China (Grant No. U20A20223). This work is supported by A*STAR through the Industry Alignment Fund - Industry Collaboration Projects Grant.

References

- [1] Neuroscience: why do we see faces in everyday objects? <https://www.bbc.com/future/article/20140730-why-do-we-see-faces-in-objects>.
- [2] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. Bringing portraits to life. *TOG*, 36(6):1–13, 2017.
- [3] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.
- [4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999.
- [5] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017.
- [6] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *CVPR*, 2020.
- [7] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, 2019.
- [8] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.
- [9] Gerald Farin, Josef Hoschek, and M-S Kim. *Handbook of computer aided geometric design*. Elsevier, 2002.
- [10] Chaoyou Fu, Yibo Hu, Xiang Wu, Guoli Wang, Qian Zhang, and Ran He. High-fidelity face manipulation with extreme poses and expressions. *TIFS*, 16:2218–2231, 2021.
- [11] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. Dual variational generation for low shot heterogeneous face recognition. In *NeurIPS*, 2019.
- [12] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. Dvg-face: Dual variational generation for heterogeneous face recognition. *TPAMI*, 2021.
- [13] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. Automatic face reenactment. In *CVPR*, 2014.
- [14] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *TOG*, 37(6):1–12, 2018.
- [15] Andrew S Glassner. *Graphics gems*. Elsevier, 2013.
- [16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.
- [17] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, 2020.
- [18] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *AAAI*, 2020.
- [19] WANG Hao and YANG Zhigang. Face pareidolia and its neural mechanism. *Advances in Psychological Science*, 26(11):1952–1960, 2018.
- [20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.
- [21] Yibo Hu, Xiang Wu, Bing Yu, Ran He, and Zhenan Sun. Pose-guided photorealistic face rotation. In *CVPR*, 2018.
- [22] Po-Hsiang Huang, Fu-En Yang, and Yu-Chiang Frank Wang. Learning identity-invariant motion representations for cross-id face reenactment. In *CVPR*, 2020.
- [23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.
- [24] Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. Neural style-preserving visual dubbing. *TOG*, 38(6):1–13, 2019.
- [25] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *TOG*, 37(4):1–14, 2018.
- [26] A. Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *CACM*, 2017.
- [27] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. In *CVPR*, 2019.
- [28] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *SIGGRAPH Asia*, 36(6):194:1–194:17, 2017.
- [29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [30] Chang Liu, Dezhao Luo, Yifei Zhang, Wei Ke, Fang Wan, and Qixiang Ye. Parametric skeleton generation via gaussian mixture models. In *CVPRW*, 2019.
- [31] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018.
- [32] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, 2019.
- [33] Shengju Qian, Kwan-Yee Lin, Wayne Wu, Yangxiaokang Liu, Quan Wang, Fumin Shen, Chen Qian, and Ran He. Make a face: Towards arbitrary high fidelity face manipulation. In *ICCV*, 2019.
- [34] Geetha Ramachandran. A combined distance measure for 2d shape matching. In *International Conference on Computer Vision and Image Analysis Applications*. IEEE, 2015.
- [35] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, 2013.
- [36] David Salomon. *Curves and surfaces for computer graphics*. Springer Science & Business Media, 2007.
- [37] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-icp. In *Robotics: science and systems*, 2009.
- [38] Eugene V Shikin and Alexander I Plis. *Handbook on Splines for the User*. CRC press, 1995.

- [39] K. Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2014.
- [40] Linsen Song, Jie Cao, Lingxiao Song, Yibo Hu, and Ran He. Geometry-aware face completion and editing. In *AAAI*, 2019.
- [41] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *arXiv preprint arXiv:2001.05201*, 2020.
- [42] Yi-Zhe Song. Béziersketch: A generative model for scalable vector sketches. In *ECCV*, 2020.
- [43] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *TOG*, 36(4):1–13, 2017.
- [44] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*, 2020.
- [45] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *TOG*, 34(6):183–1, 2015.
- [46] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016.
- [47] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Facevr: Real-time gaze-aware facial reenactment in virtual reality. *TOG*, 37(2):1–15, 2018.
- [48] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017.
- [49] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020.
- [50] Susan G. Wardle, Jessica Taubert, Lina Teichmann, and Chris I. Baker. Rapid and dynamic processing of face pareidolia in the human brain. *Nature Communications*, Nat Commun(11), 2020.
- [51] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *CVPR*, 2019.
- [52] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, 2018.
- [53] Zili Yi, Qiang Tang, Vishnu Sanjay Ramiya Srinivasan, and Zhan Xu. Animating through warping: An efficient method for high-quality facial expression animation. In *ACM MM*, 2020.
- [54] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019.
- [55] Juyong Zhang, Keyu Chen, and Jianmin Zheng. Facial expression retargeting from human to avatar made easy. *TVCG*, 2020.
- [56] Yang Zhou, DIngzeyu Li, Xintong Han, Evangelos Kalogerakis, Eli Shechtman, and Jose Echevarria. Makeittalk: speaker-aware talking head animation. In *SIGGRAPH Asia*, 2020.