POCE: Pose-Controllable Expression Editing

Rongliang Wu, Yingchen Yu, Fangneng Zhan, Jiahui Zhang, Shengcai Liao, *Senior Member, IEEE*, Shijian Lu*

***Abstract*—Facial expression editing has attracted increasing attention with the advance of deep neural networks in recent years. However, most existing methods suffer from compromised editing fidelity and limited usability as they either ignore pose variations (unrealistic editing) or require paired training data (not easy to collect) for pose controls. This paper presents POCE, an innovative pose-controllable expression editing network that can generate realistic facial expressions and head poses simultaneously with just unpaired training images. POCE achieves the more accessible and realistic pose-controllable expression editing by mapping face images into UV space, where facial expressions and head poses can be disentangled and edited separately. POCE has two novel designs. The first is self-supervised UV completion that allows to complete UV maps sampled under different head poses, which often suffer from self-occlusions and missing facial texture. The second is weakly-supervised UV editing that allows to generate new facial expressions with minimal modification of facial identity, where the synthesized expression could be controlled by either an expression label or directly transplanted from a reference UV map via feature transfer. Extensive experiments show that POCE can learn from unpaired face images effectively, and the learned model can generate realistic and high-fidelity facial expressions under various new poses.**

***Index Terms*—Facial Expression Editing, Image Synthesis, Generative Adversarial Networks.**

I. Introduction

Facial expression editing aims to edit the expression of a face image without changing the face identity. Automated and realistic expression editing has attracted increasing interest due to its wide range of applications in photography, animation, etc. However, generating high-fidelity expressions is a challenging task as the human visual system is very sensitive to tiny changes in facial expressions [1], [2]. While considering concurrent variations in facial expressions and head poses as in practical situations, realistic and high-fidelity facial expression editing becomes even more challenging.

Automated facial expression editing has achieved quite impressive progress in recent years. One typical approach is pose-fixed editing that focuses on expression editing only without handling the head pose of the edited face image. Leveraging the recent development of generative adversarial networks (GANs) [3], several studies [4]–[7] formulate pose-fixed expression editing as an unpaired image-to-image translation task and require just a single source image for inference. However, these methods suffer from *limited realism* as head poses and facial expressions usually vary simultaneously by nature. Additionally, they require the edited face image to be frontal or almost frontal and cannot handle many face images that are under non-frontal poses [4]–[7].

Pose-controllable expression editing aims to edit facial expressions and head poses simultaneously with minimal modification of facial identity features. It has been attracting increasing interest from both academia and industry since it is better aligned with natural expression changes. Most existing studies exploit 3D facial structures as extracted by 3D modeling [8] or deep generative networks [3], but suffer from two typical constraints. First, they require paired images (i.e., face images of the same person with different expressions and poses) [9], [10] or video sequences [2], [11]–[13] for training, which are not easy to collect in practice and accordingly limit the *usability* of these methods greatly. Second, they usually condition on facial landmarks that inherently encode facial expressions, facial identity and head poses altogether [2], [10], [12]. The editing of such highly entangled expressions and poses in landmarks tends to introduce undesired modification of face identity (e.g., face shape), and this degrades the editing *flexibility* and editing *quality* greatly.

This paper presents a novel **PO**se-**C**ontrollable **E**xpression editing (POCE) network that can edit facial expressions and head poses simultaneously with just unpaired training images. Inspired by the idea of UV maps that project 3D texture to a 2D pose-invariant template with universal per-pixel alignment, POCE converts face images into UV maps and disentangles expression editing and pose generation elegantly. Given a face image, we fit a 3D face model to sample a facial UV map which allows to edit expressions in the UV space and render the edited UV to new poses accurately. POCE has two novel designs that enable pose-controllable expression editing. The first is self-supervised UV completion that allows to generate complete UV texture from face images of different poses with various self-occlusions. The second is weakly supervised UV editing that allows to generate realistic expressions with minimal modification of face identity, where the synthesized expression could be controlled by either an expression label or directly transplanted from a reference UV map via feature transfer. Extensive experiments show that POCE can achieve realistic pose-controllable expression editing with just unpaired training data.

The contributions of this work are threefold. *First*, we propose POCE, an innovative network that can edit facial expressions and head poses simultaneously with just unpaired training images. *Second*, we introduce UV maps for pose-

R. Wu is with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. Partial work done during study at Nanyang Technological University, Singapore.

Y. Yu, J. Zhang, and S. Lu are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore.

F. Zhan is with the Nanyang Technological University, Singapore and Max Planck Institute for Informatics, Germany.

S. Liao is with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes more details and experimental results of the proposed method. Contact ronglian001@e.ntu.edu.sg for further questions about this work.

* indicates the corresponding author. Email: shijian.lu@ntu.edu.sg

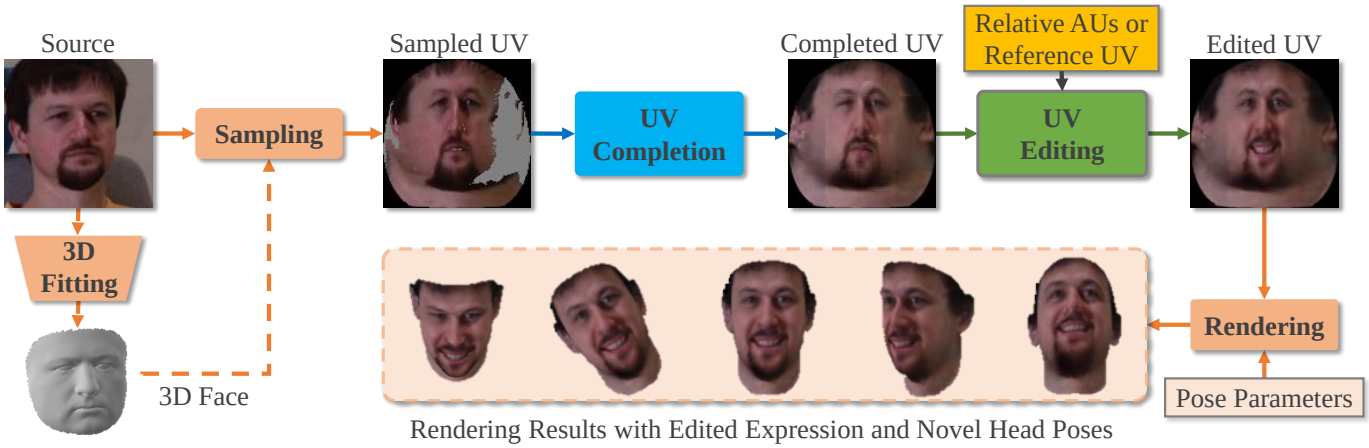


Fig. 1. The framework of the proposed POCE: Given a *Source* face, we first fit a 3D face model to the face image to convert it to a *Sampled UV* where expression editing and poses generation can be disentangled conveniently. The *Sampled UV* is then fed to a *UV Completion* module and a *UV Editing* module which complete the UV map and edit the expression conditioned on *Relative AUs* or transferred from *Reference UV* via feature transfer, respectively. Finally, the *Edited UV* is rendered to novel head poses conditioned on *Pose Parameters* for generating realistic expressions and poses simultaneously.

controllable facial expression editing. On top of that, we design novel UV completing and UV editing techniques which achieve UV completion of self-occluded face images and identity-preservative expression editing, respectively. *Third*, extensive experiments show that POCE achieves superior pose-controllable expression editing quantitatively and qualitatively.

II. Related Work

A. Facial Expression Editing

Automated expression editing has been studied for years and most existing works can be grouped into 3D model based methods and generation based methods.

3D Model based Methods: Classical expression editing methods model 3D face structures with 3D Morphable Models (3DMMs). For example, [8] presents the first public 3DMM, where linear model was created to represent face variations. [14] introduces a multi-linear model to map one person's performance to facial animations of another. [15] designs Face2Face for expression tracking and re-targeting. 3DMMs can jointly model expressions and poses, but they tend to produce blurs due to the Gaussian assumption [16]. In addition, they require hard-to-collect 3D face scans or videos in training, which limits their usability greatly.

Generation based Methods: Generation based methods exploit deep generative networks [3] for facial expression editing. For example, [4] proposes StarGAN for multi-modality editing conditioned on discrete expression labels. [5] designs GANimation for continuous expression editing. [6] presents Cascade EF-GAN for progressive expression editing. [17] introduces LEED for label-free expression editing. [7] proposes MSF for fine-grained expression editing. These prior studies can work with unpaired images, but they can only handle pose-fixed editing which degrades the editing realism greatly as expressions and poses usually vary concurrently by nature.

Pose-controllable editing via deep generation has attracted increasing interest recently. Due to the lack of 3D face structures, most existing works [2], [10]–[12], [18]–[22] require paired expression images or video sequences in training, which impairs their usability greatly. In addition, they usually condition on facial landmarks [2], [10], [12], [18], [19], [21] that naturally entangle expressions, identity and poses rigidly. Several works [11], [13], [20] attempt to use predicable latent features but they work on entangled expressions and poses. Recently, [23] introduces DPE, which aims to learn disentangled pose and expression representations from video data for portrait editing.

The proposed POCE requires only unpaired images in training but can edit facial expressions and head poses simultaneously. Besides, it converts face images into UV maps where facial expression editing and head pose generation can be achieved independently, which improves editing flexibility and controllability greatly.

B. Image Completion

Image completion aims at filling missing pixels in images. [24] presents Context Encoder to address hole-filling problems. [25] introduces multi-scale neural patch synthesis for preserving contextual structures with high-frequency details. [26] designs gated convolution to complete irregular holes. [27] proposes UV-GAN to complete self-occluded UV maps but requires large-scale incomplete/complete UV pairs (expensive and time-consuming to collect) in network training. [28] introduces OSTeC for iterative texture completion, which needs to be optimized for each image in inference.

The proposed UV completion network completes self-occluded UV maps in a self-supervised manner without requiring ground-truth UV map. In addition, the trained model can be applied to images collected from different people, which makes it accessible and scalable to different users and tasks.

III. Proposed Method

A. Overview

Fig. 1 shows the POCE pipeline. Inspired by the idea that UV maps project 3D texture data to a 2D pose-invariant

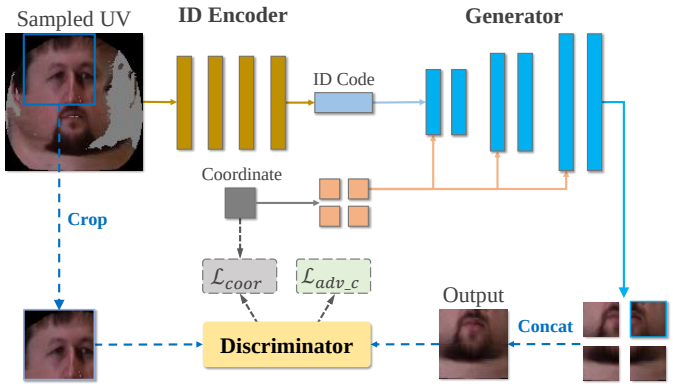


Fig. 2. Illustration of the proposed patch-based UV completion network: Given a *Sampled UV* as input, an *ID Encoder* first extracts the face identity information. A *Generator* then takes the extracted *ID Code* and the randomly sampled macro *Coordinate* as input to generate a complete macro UV patch. Specifically, the *Generator* first generates consecutive micro patches with the automatically-derived micro coordinates and stitches them to obtain the output macro patch. The real and generated macro UV patches are then forwarded to the *Discriminator* for adversarial learning.

template, we convert a face image into a UV map where expression editing and head pose generation can be disentangled elegantly. Given a *Source* image, we first fit a 3D face model to sample a facial UV map which often contains certain missing regions due to self-occlusions. The *Sampled UV* is then fed to the proposed *UV Completion* which generates missing texture to produce a *Completed UV*. The *Completed UV* is further fed to the proposed *UV Editing* that generates *Edited UV* with target expression conditioned on *Relative AUs* or transplanted from *Reference UV*. With the head *Pose Parameters*, the completed and edited UV map is finally rendered to a target head pose to achieve pose-controllable expression editing. More details of *UV Completion*, *UV Editing*, *3D Face Fitting* and *Rendering* will be discussed in the ensuing subsections.

B. UV Completion

Due to self-occlusions, the facial UV map sampled from face image is usually incomplete with missing texture, which affects both realistic expression editing and new poses generation. We design a UV completion technique that learns to complete UV map in a patch-based manner, which can generate complete UV maps from incomplete ones and provide realistic texture for face rendering under various new poses.

1) Network Overview: As illustrated in Fig. 2, our patch-based UV completion network consists of an identity encoder E_I, a generator G_c and a discriminator D_c. Given a sampled UV map, E_I first extracts identity information which is shared among all patches of the same face image. G_c then generates a complete macro UV patch based on the extracted identity information and a randomly sampled macro coordinate. Specifically, G_c first generates consecutive micro patches with the micro coordinates derived from the sampled macro coordinate and stitches them to obtain the macro patch. D_c aims to distinguish the generated macro patches against real ones (that cropped from incomplete UV maps) and it guides the generator to synthesize coherent contents for adjacent

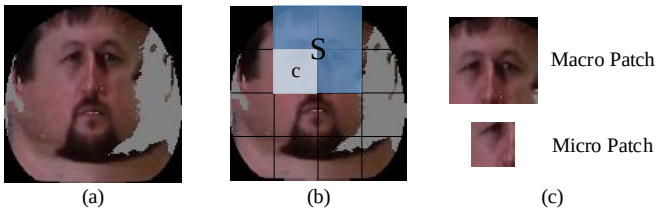


Fig. 3. Illustrations of the coordinate systems in the proposed UV completion: (a) shows a sampled UV, (b) illustrates the macro (s) and micro (c) coordinate systems used in our design, and (c) illustrates the macro and micro patches used for training the UV completion network.

micro patches in an adversarial manner. Once the model is trained, it synthesizes patches at all coordinates and stitches them as a completed UV map.

2) Patch-based UV Completion: Most existing UV completion methods are trained with paired UV including a sampled incomplete UV map and a complete UV map (as the ground truth in supervised training) of the same person [27], but capturing complete face UV has various restrictions in equipment (requiring 3D scanner or multi-view cameras), portraiture light, etc. To the best of our knowledge, only one public dataset provides complete UV [27], but the UV maps are collected in lab environments, which do not generalize well to face images in the wild. We design a patch-based UV completion network that can complete UV maps in a self-supervised manner without requiring complete ground-truth UV. Our design is inspired by the observations that UV maps are highly aligned (with universal per-pixel alignment) in the 2D pose-invariant template and the sampled UV is composed of visible texture in a continuous region. We can thus crop complete UV patches from different locations of incomplete UVs though we do not have complete UV in training. With the cropped patches, our UV completion learns to complete UV in a patch-based manner conditioning on the spatial coordinates and face identity as extracted from incomplete UV.

3) Coordinate System: The proposed UV completion module learns to complete facial UV maps in a patch-based manner without requiring paired UV data for network training. The trained UV completion model can synthesize patches at all coordinates and stitch them to obtain a completed UV map. A coordinate system is thus required for providing spatial guidance. Our idea of designing the coordinate system is straightforward: it divides a UV map into multiple non-overlapping patches and assigns a unique coordinate to each patch to encode the spatial location information.

However, directly generating patches and stitching them together often leads to artifacts around the patch boundaries and affects realistic UV completion. Inspired by [29], we address this issue by training the network to generate consecutive micro patches and stitch them to form a macro patch, and introducing an adversarial loss to penalize the incoherent contents within the stitched macro patch. This strategy encourages the network to generate seamless boundaries for adjacent micro patches, which helps suppress the artifacts effectively. To this end, we design micro and macro coordinate systems for handling patches of different sizes as illustrated in Fig. 3.

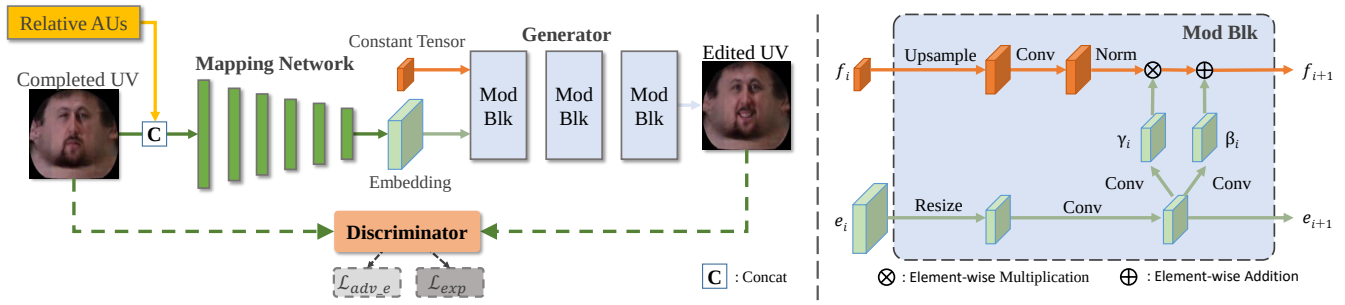


Fig. 4. Illustration of the proposed UV editing network: The *Mapping Network* first transforms the *Completed UV* and *Relative AUs* to the latent *Embedding* that encodes source identity attribute and desired facial expression. The *Generator* then takes as input the *Embedding* and the *Constant Tensor* that encodes the coarse geometry prior shared among all UV maps to produce *Edited UV*. Detailed structure of the modulation block (*Mod Blk*) is shown on the right.

Specifically, the micro coordinate system divides a UV map into m by n micro patches without overlapping:

$$C = \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,n} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m,1} & c_{m,2} & \cdots & c_{m,n} \end{pmatrix},$$

while the macro coordinate system is defined as:

$$S = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,n'} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,n'} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m',1} & s_{m',2} & \cdots & s_{m',n'} \end{pmatrix},$$

where $s_{i,j} = [c_{i:i+p-1,j:j+q-1}]$, indicating that the macro patch is formed by $p \times q$ consecutive micro patches. p and q denote the number of micro patches that construct a macro patch in horizontal and vertical directions, respectively. With the given notations, we can easily derive $m' = m - p + 1$, $n' = n - q + 1$.

We crop all macro UV patches from each incomplete UV in a raster-scan order according to the macro coordinate system, and empirically treat patches that contain more than 95% valid pixels as complete UV patches.

4) Loss Functions: The loss for training the UV completion network consists of four terms:

$$\mathcal{L}_C = \mathcal{L}_{adv\_c} + \lambda_{coor}\mathcal{L}_{coor} + \lambda_{cont}\mathcal{L}_{cont} + \lambda_{sym}\mathcal{L}_{sym}, \quad (1)$$

where the weights λs are used to balance each component.

The first term is a patch-level adversarial loss that guides the discriminator to distinguish real/generated UV patches and the generator to generate complete and realistic patches:

$$\begin{aligned} \mathcal{L}_{adv\_c} =& \mathbb{E}[D_c(\mathbf{T}_G^i)] - \mathbb{E}[D_c(\mathbf{T}_R^j)] \\ &+ \lambda_c \mathbb{E}[(\|\nabla D_c(\hat{\mathbf{T}})\|_2 - 1)^2], \end{aligned} \quad (2)$$

where $\mathbf{T}_G^i$ is the generated UV patch with coordinates s^i, $\mathbf{T}_R^j$ is the real complete UV patch with coordinates s^j cropped from the visible region, $\hat{\mathbf{T}}$ is the interpolated data introduced in [30] and λ_c is the balancing weight.

The second term is the coordinate regression loss that guides the generator to align the generated patches with the given spatial coordinates. Specifically, we add an auxiliary head on top of the discriminator to predict spatial coordinate, and apply L2 loss on the coordinates of both generated and real patches:

$$\mathcal{L}_{coor} = \mathbb{E}[\|D_x(\mathbf{T}_G^i) - s^i\|_2^2] + \mathbb{E}[\|D_x(\mathbf{T}_R^j) - s^j\|_2^2], \quad (3)$$

where D_x is the coordinate regression head on top of D_c.

The third term is the content loss that penalizes the difference between the generated UV and the sampled UV for the regions with valid texture. It encourages the network to retain the valid UV texture as faithfully as possible while generating missing texture, and helps preserve the identity information of the sampled UV. Specifically, we penalize L1 difference between the generated patch and the one cropped from UV_{sam} at the same location (for the regions with valid texture):

$$\mathcal{L}_{cont} = \mathbb{E}\|(\mathbf{T}_G^i - \mathbf{T}_R^i) \otimes M^i\|_1, \quad (4)$$

where $\otimes$ denotes element-wise multiplication, M^i is a binary mask that indicates the validity of the texture within the sampled UV patch. The obtaining process of M^i will be introduced in Section III.D.

The last term is the symmetry loss that exploits facial symmetry to guide the network to generate missing texture:

$$\mathcal{L}_{sym} = \|(\mathbf{T}_G^i - flip(\mathbf{T}_G^{-i}))\|_1, \quad (5)$$

where $flip(\mathbf{T}_G^{-i})$ is the horizontal flipped patch (at mirror location) of $\mathbf{T}_G^i$.

C. UV Editing

With a completed UV map, the proposed UV editing network aims to edit its expression with minimal modification of facial identity features. We employ the widely adopted Facial Action Coding System (FACS) [31] to describe facial expressions in terms of the intensities of continuous Action Units (AUs). Specifically, we exploit relative AUs as expression conditions for training the UV editing network and design a modulation-based generator that incorporates spatially-varying modulation to edit the facial expressions. Once the model is trained, it allows to edit facial expressions that can be either controlled by the relative AUs that encode desired expression information or directly transplanted from a reference UV map via feature transfer.

1) Network Overview: As illustrated in Fig. 4, our UV editing network consists of a mapping network M_e, a generator G_e and a discriminator D_e. M_e takes the completed UV map (by UV completion network) and relative AUs as input, and maps them to the latent embedding that encodes the identity attribute of source UV map as well as the desired facial expression. G_e then takes the embedding and a constant tensor that encodes the coarse geometry prior that is shared among all UV maps as input to generate an edited UV with target expression. D_e evaluates the photo-realism of edited UV and examines whether it contains desired expression information.

2) Relative AUs: Inspired by [7], [32] that use difference vector to control image attributes, we train the UV editing network conditioning on relative AUs (AU_{rel}), which are defined as the difference between the AUs of source image (AU_{src}) and target AUs (AU_{tgt}) that encode desired expressions:

$$AU_{rel} = AU_{tgt} - AU_{src}. \tag{6}$$

Training the UV editing network conditioning on relative AUs has three benefits. First, relative AUs can better guide the network to focus on interested regions as compared with absolute AUs. Existing works [5], [6] feed absolute AUs and a source image to the network which first predicts an attention map to identify regions-of-interest and then performs editing. This requires the network to implicitly estimate source AUs and compare them with target AUs to generate the attention map. In contrast, utilizing relative AUs allows the residual information to be explicitly injected into the network, which guides the network to focus on the interested regions and makes it converge faster. Second, using relative AUs helps generate more accurate editing especially when only specific facial regions require editing. To edit specific regions, models using absolute AUs need to estimate the corresponding AUs in source face, modify their intensities and generate edited expression with the modified AUs. However, the AU estimation may suffer from errors, which leads to editing of unrelated facial attributes. With relative AUs, models just need to modify relative AUs intensities of interested regions and set the intensities of remaining AUs to zero. This mitigates undesired AU manipulations and leads to more accurate editing. Third, editing facial expressions conditioning on relative AUs facilitates the trained model to edit expressions by directly transplanting the expression from the reference UV map to the source UV via feature transfer (more details to be shared in Section III-C4).

3) Modulation-based Generator: Most existing facial expression editing models [4]–[6] employ encoder-decoder architecture to generate the edited output. They first leverage an encoder to transform the source images to high-level representations, and then forward them to a decoder to produce the editing results. However, as discussed in [32], the encoder-decoder architecture tends to discard fine details of source images, leading to blurry editing results. In addition, the learned representations are unstructured which do not support facial expression editing via feature transfer (i.e., replacing partial representations of the source image with that of the reference image to achieve expression editing).

To mitigate the above issues, we design a modulation-based generator that incorporates spatially-varying modulation [33]–[35] for expression editing in UV maps. The idea of designing the modulation-based generator is inspired by the observation that UV maps project 3D facial texture to a 2D pose-invariant template with universal per-pixel alignment, i.e., the facial features of different people (e.g., eyes, nose and mouth) are projected to similar location in the UV map. Hence, the UV maps of different people with different expressions share similar underlying geometry but vary in detailed texture information only. With this observation, the proposed modulation-based generator takes a constant tensor $t \in \mathbb{R}^{64\times8\times8}$ as initial input to encode the coarse geometry prior that is shared among all facial UV maps. It then forwards t to multiple modulation blocks, which gradually produce feature maps of higher spatial resolution and inject detailed texture information into the feature maps (via spatially-varying modulation). Finally, the generator transforms the output feature maps of the last modulation block to the edited UV that combines the desired facial expression and the identity attribute of the source UV. Since the generator is differentiable in its input, we follow [36] to optimize t together with the network weights via backpropagation in network training, aiming to find the optimal t that effectively encodes the coarse geometry shared among all UV maps. Once t is searched, it is frozen and used as a fixed input (or "constant") to the expression editing network for editing across different faces. Note the method in [36] focuses on learning a 1D pose vector for transforming unaligned images into a canonical view. We instead optimize a 3D tensor to encode the coarse geometry (shared among all UV maps) to facilitate expression editing.

The modulation block modulates the feature maps with spatially-varying modulation parameters. These parameters are learned from the latent embedding (produced by a mapping network) that encodes the desired facial expression as well as the identity attribute of the source UV map. The modulation operation of the i-th modulation block could be formulated as:

$$f_{i+1} = (\gamma_i \otimes \frac{f_i - \mu_i}{\sigma_i}) \oplus \beta_i, \tag{7}$$

where f_i is the input feature maps ($f_0 = t$), μ_i and σ_i are the mean and standard deviation of f_i, γ_i and β_i are the modulation parameters which have the same size as f_i, $\otimes$ and $\oplus$ denote the element-wise multiplication and addition operation, respectively.

With the modulation blocks, the designed generator can generate much sharper editing results than most existing methods [4]–[6] that employ an encoder-decoder architecture. In addition, the constant tensor and element-wise modulation explicitly help the generator build spatial connection between the modulated parameters and the output UV map. As a result, partial change in the modulated parameters will lead to local editing on the generated UV, which enables to edit facial expression via feature transfer.

4) Expression Editing via Feature Transfer: In addition to edit expressions conditioning on relative AUs, the trained UV editing network allows to directly transplant expression from reference UV to source UV via feature transfer with respective

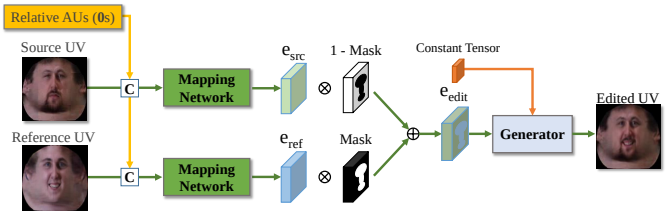


Fig. 5. Illustration of the proposed expression editing via feature transfer pipeline: The *Source UV* and *Reference UV* are first fed to the well-trained *Mapping Network* together with the *Zeroized Relative AUs* (where all elements are set to 0) to retrieve the corresponding embeddings (e_{src} and e_{ref}) that encode the identity and expression of the original UV maps. e_{src} is then blended with e_{ref} with respective to a given *Mask* to transfer the expression-related features from e_{ref} to e_{src}. Finally, the edited embedding e_{edit} is forwarded to the trained *Generator* together with the *Constant Tensor* to produce the *Edited UV* that combines the facial expression of the reference UV and the identity attribute of the source UV.

to a given mask. Note that the mask could be in any shape, which allows to simultaneously transfer expression around eyes, nose and mouth regions (that contain most expression-related information [6], [37], [38]), or just partial expression around interested facial region (e.g., eyes only or mouth only). In this way, we can flexibly transfer the desired expression from reference to source UV without requiring any expression annotation, which improves the editing flexibility greatly.

Fig. 5 shows the expression editing via feature transfer pipeline. Given the source UV and the reference UV that provide identity and desired expression information, we first forward them to the trained mapping network together with the zeroized relative AUs (where all elements are set to 0) to retrieve the corresponding embeddings (e_{src} and e_{ref}) that encode the identity and expression information of the original UV. We then apply alpha blending to e_{src} and e_{ref} to obtain the edited embedding e_{edit}:

$$e_{edit} = m \otimes e_{ref} \oplus (1 - m) \otimes e_{src}, \tag{8}$$

where m is a binary mask that is resized by max pooling to match the spatial resolution of the embeddings, $\otimes$ and $\oplus$ are the same as Eq. (7). As discussed in Section III-C3, the embedding is mapped to spatially-varying modulation parameters which inject detailed texture information into the feature maps in an element-wise manipulation manner. Transferring the reference embedding e_{ref} around eyes, nose and mouth regions to that of e_{src} thus transfers most expression information from the reference UV to the source UV. At the other end, partial embedding transfer leads to local expression editing, where the edited regions can be specified by masks as illustrated in Fig. 7. Note we perform feature transfer on the modulation parameters which have larger spatial resolution (than the embedding) to achieve fine-grained manipulation in the experiments, but we explain the transfer operation on the embedding for simplicity.

Finally, we forward the edited embedding e_{edit} as well as the constant tensor t to the trained generator to produce the edited UV map that combines the identity attribute of source UV and the expression attribute of reference UV.

5) Loss Functions: The objective function for training the UV editing network consists of three terms:

$$\mathcal{L}_E = \mathcal{L}_{adv\_e} + \lambda_{exp}\mathcal{L}_{exp} + \lambda_{cyc}\mathcal{L}_{cyc}. \tag{9}$$

The first term is an adversarial loss [30] for improving the photo-realism of the edited UV:

$$\begin{aligned}\mathcal{L}_{adv\_e} =& \mathbb{E}[D_e(UV_{src})] - \mathbb{E}[D_e(UV_{edit})] \\ &+ \lambda_e \mathbb{E}[(\|\nabla D_e(\widehat{UV})\|_2 - 1)^2],\end{aligned} \tag{10}$$

$$UV_{edit} = G_e(M_e(UV_{src}, AU_{rel}), t), \tag{11}$$

where UV_{src} is the completed source UV map (by UV completion network), UV_{edit} is the output of UV editing network, $\widehat{UV}$ is the interpolated data introduced in [30] and λ_e is the balancing weight, respectively.

The second term is a conditional expression loss that guides the generator to generate a UV map with desired expression:

$$\begin{aligned}\mathcal{L}_{exp} =& \mathbb{E}[\|D_y(UV_{src}) - AU_{src}\|_2^2] \\ &+ \mathbb{E}[\|D_y(UV_{edit}) - AU_{tgt}\|_2^2],\end{aligned} \tag{12}$$

where D_y is the AUs regression head on top of D_e.

The third term is a cycle reconstruction loss that guides the generator to keep facial identity and personal attributes of the source UV after editing:

$$\begin{aligned}\mathcal{L}_{cyc} =& \mathbb{E}\|UV_{src} - G_e(M_e(UV_{src}, 0), t)\|_1 + \\ & \mathbb{E}\|UV_{src} - G_e(M_e(UV_{edit}, -AU_{rel}), t)\|_1.\end{aligned} \tag{13}$$

D. 3D Face Fitting and Rendering

1) 3D Face Fitting: We fit a pre-trained 3DMM to face images to derive facial UV maps, which serves as a prerequisite step for our method. A number of open-source 3DMM models are available [39]–[41] and we adopt 3DDFA [39] in our experiments, which is lightweight and capable of real-time image processing. Specifically, we apply 3DDFA to estimate 3D face shape $\mathcal{S}$ and head pose $\mathcal{P}$ from the face image I:

$$\{\mathcal{S}, \mathcal{P}\} = Fitting(I), \tag{14}$$

where $\mathcal{P}$ is parameterized by Euler's angles (*pitch*, *yaw*, *roll*), translation and a face scale factor. The 3D face is then projected to the image plane with weak perspective projection and the facial UV map UV_{sam} can be sampled from I with the standard rasterization pipeline [42]:

$$\{UV_{sam}, M\} = \mathbf{F}(I, \mathcal{V}(\mathcal{S}, \mathcal{P})), \tag{15}$$

where $\mathbf{F}$ denotes the sampling operation, $\mathcal{V}$ denotes the projection operation and M is a binary mask indicating the validity of the texture within the sampled UV, which can be obtained by employing depth buffer-based method [39].

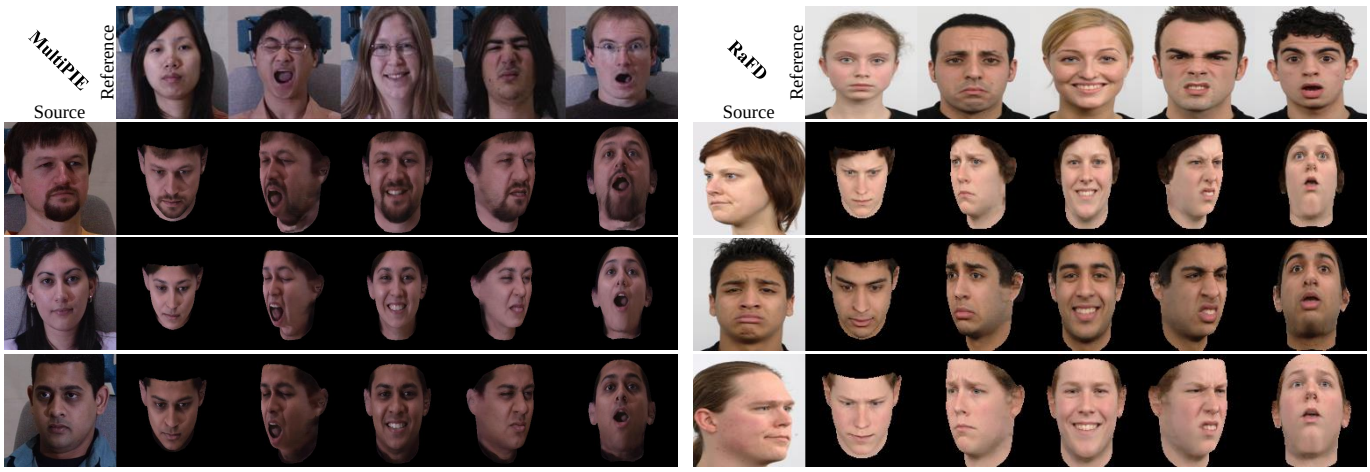


Fig. 6. Expression and pose editing by the proposed POCE over MultiPIE and RaFD: In each sub-figure, the first column shows the source images, and the first row shows the reference images with target expressions. The rest rows and columns show our editing. POCE is capable of editing facial expressions and head poses realistically and simultaneously in a disentangled manner.

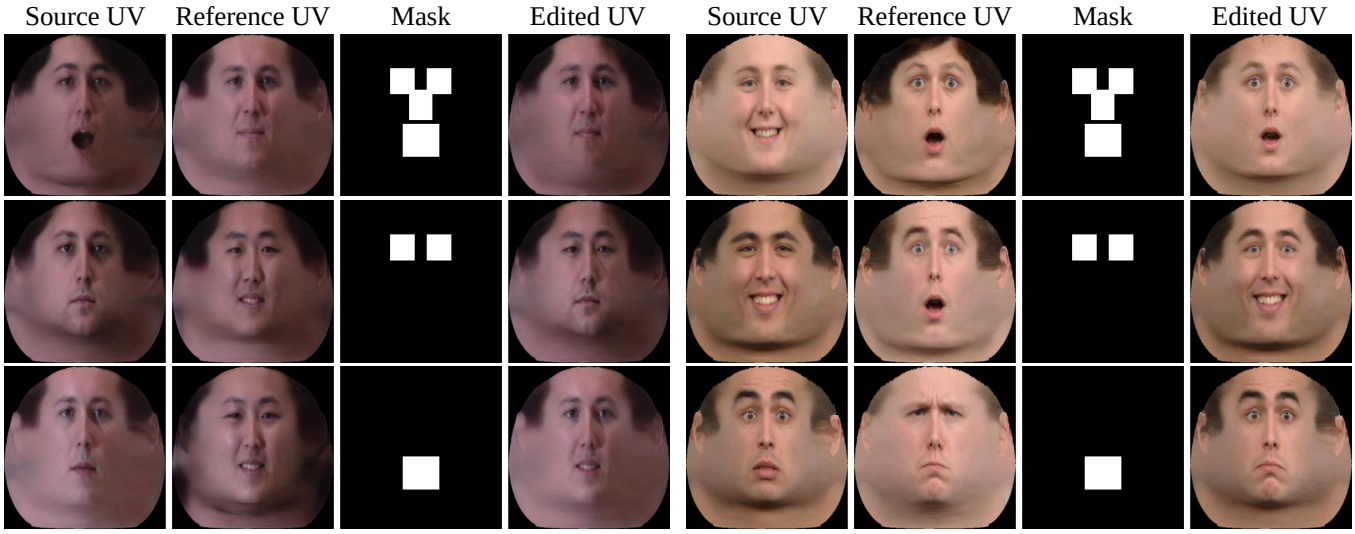


Fig. 7. Facial expression editing via feature transfer by POCE over MultiPIE and RaFD: Given a mask, POCE can transfer expression information around interested facial regions from reference UV to source UV and achieve editing of partial expression effectively. In each sub-figure, rows 1, 2 and 3 show the expression editing around eyes+nose+mouth, eyes only, and mouth only, respectively.

2) Rendering: The sampled UV map is fed to our UV completion and UV editing modules for generating complete texture and editing expression, respectively. With a completed and edited UV map UV_{edit}, we can render it to target pose $\mathcal{P}_{tgt}$ (which could be either tuned by user or estimated from a reference image) with an off-the-shelf renderer to achieve pose-controllable expression editing:

$$I_{output} = \mathbf{R}(UV_{edit}, \mathcal{S}, \mathcal{P}_{tgt}), \tag{16}$$

where $\mathbf{R}$ denotes the renderer. In our implementation, we use an open-sourced renderer [43] to perform rendering without any training.

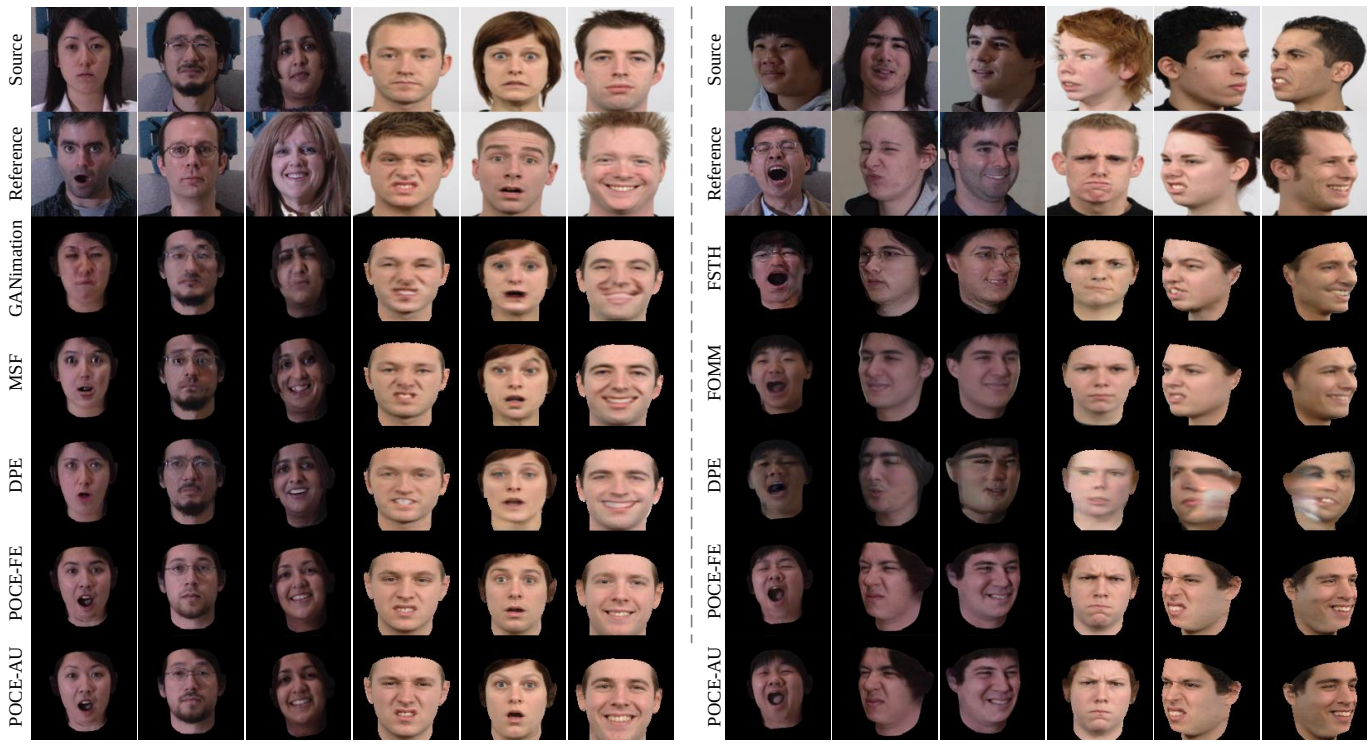
IV. Experiments

A. Settings

Datasets: The proposed POCE is evaluated over datasets MultiPIE [44] and Radboud Faces (RaFD) [45]. MultiPIE consists of more than 750,000 images of 337 identities showing different facial expressions. RaFD contains 8,040 facial expression images of 67 participants collected from different viewpoints. We randomly sample 90% images for training and the rest for testing for both datasets.

Evaluation Metrics: We perform quantitative evaluations and comparisons with several widely adopted metrics as follows:

• Identity Embedding Distance (IED). IED measures L2 distances between the embedded features of source and edited faces that are extracted by a pre-trained face recognition



(a) Pose-fixed Expression Editing (b) Pose-controllable Expression Editing

Fig. 8. Expression editing by POCE and the state-of-the-art methods over MultiPIE and RaFD: In each sub-figure, columns 1-3 show the editing over MultiPIE images while columns 4-6 show the editing over RaFD images. POCE-FE and POCE-AU indicate the POCE editing results generated via feature transfer and conditioning on relative AUs, respectively. It can be seen that POCE can produce more realistic editing with better details and less artifacts than the state-of-the-art methods in both pose-fixed and pose-controllable expression editing.

model [46]. A lower IED indicates better identity preservation of the edited face images.

• Expression Distance (EXD). EXD measures L2 distances between the AUs intensities of reference and edited faces estimated by OpenFace [47]. Lower EXDs mean higher similarity between the expressions of the edited and reference faces.

• Fréchet Inception Distance (FID) [48] and Inception Score (IS) [49]. FID and IS are computed based on the extracted features of pre-trained models [50], [51]. Lower FID and higher IS indicate better image quality of the edited faces.

• Subjective evaluations. We also conducted Amazon-Mechanical-Turk (AMT) user studies to evaluate the perceptual realism of the edited images. Specifically, the subjects are tasked to evaluate the editing quality based on different criteria. The first is Realism Assessment (RA), where the subjects are presented with real and edited images and tasked to tell whether the images are real or fake by providing their ratings from 1 to 5. The second is Expression Similarity (ES), where the subjects are presented with a pair of images that consist of a reference image and an edited image (by different methods) and tasked to evaluate whether they contain similar facial expression by providing their ratings from 1 to 5. The third is User Preference (UP), where the subjects are presented with a reference image and a set of randomly-ordered images edited by different methods, and tasked to identify the most realistic one. We report mean opinion score of the ratings collected from the AMT users, where larger scores indicates better perceptual quality of the edited images.

Implementation Details: The UV completion and UV editing networks are trained separately. Specifically, we first train the UV completion network with the sampled UV until it converges. Then we use the completed UV maps (generated by the well-trained UV completion network) to train the UV editing network. The training is conducted on a single GeForce RTX 2080 Ti GPU with 11 GB memory and the size of UV maps is set to 256×256 in all the experiments.

• UV Completion Training Details: We use Adam optimizer [52] with $\beta_1 = 0$ and $\beta_2 = 0.999$ to optimize the parameters. We set λ_c, λ_{coor}, λ_{cont} and λ_{sym} to 100, 10, 1 and 10 to balance different losses. The size of the micro patch is set to 64×64 and 2×2 micro patches are stitched to construct a macro patch. The batch size is 2 and the total number of epochs is 30. The learning rate is 1e-4 for the first 15 epochs, then it linearly decays to 0 over another 15 epochs.

• UV Editing Training Details: We use Adam optimizer [52] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ to optimize the parameters. We set λ_e, λ_{exp} and λ_{cont} to 10, 1 and 1 to balance different losses. The batch size is set to 4. The total number of epochs is set to 100. The learning rate is set to 2e-4 for the first 50 epochs, then it linearly decays to 0 over another 50 epochs.

B. Experimental Results

Pose-Controllable Expression Editing: The proposed POCE can edit facial expressions and head poses simultaneously and

TABLE I
QUANTITATIVE COMPARISONS OF POSE-FIXED EXPRESSION EDITING BETWEEN POCE AND SOTA METHODS OVER MULTIPIE AND RAFD.

Dataset	Methods	IED ↓	EXD ↓	FID ↓	IS ↑	RA ↑	ES ↑	UP ↑
MultiPIE [44]	Real	-	-	0.00	1.578	4.49	-	-
	GANimation [5]	0.470	0.433	31.76	1.405	1.98	1.83	0.04
	MSF [7]	0.406	0.408	25.04	1.434	2.51	2.92	0.12
	DPE [23]	0.361	0.366	22.64	1.483	3.64	3.80	0.23
	POCE-FE	0.365	**0.342**	20.82	1.510	3.90	**4.36**	0.29
	POCE-AU	**0.348**	0.357	**20.43**	**1.522**	**3.97**	4.11	**0.32**
RaFD [45]	Real	-	-	0.00	1.819	4.25	-	-
	GANimation [5]	0.632	0.329	14.18	1.586	1.34	2.59	0.02
	MSF [7]	0.493	0.306	9.55	1.644	2.84	3.24	0.11
	DPE [23]	0.446	0.287	8.09	1.691	3.52	3.77	0.19
	POCE-FE	0.434	**0.274**	7.53	1.726	**4.03**	**4.45**	**0.38**
	POCE-AU	**0.427**	0.281	**7.41**	**1.737**	3.96	4.22	0.30

TABLE II
QUANTITATIVE COMPARISONS OF POSE-CONTROLLABLE EXPRESSION EDITING BETWEEN POCE AND SOTA METHODS OVER MULTIPIE AND RAFD.

Dataset	Methods	IED ↓	EXD ↓	FID ↓	IS ↑	RA ↑	ES ↑	UP ↑
MultiPIE [44]	Real	-	-	0.00	1.681	4.66	-	-
	FSTH [2]	0.872	0.518	27.04	1.499	1.50	1.94	0.05
	FOMM [13]	0.858	0.543	24.57	1.587	2.78	2.87	0.15
	DPE [23]	0.866	0.537	23.52	1.602	3.02	3.51	0.23
	POCE-FE	0.831	**0.479**	20.01	1.635	4.04	**4.20**	0.27
	POCE-AU	**0.820**	0.492	**19.46**	**1.640**	**4.13**	4.09	**0.30**
RaFD [45]	Real	-	-	0.00	1.767	4.25	-	-
	FSTH [2]	1.174	0.409	13.59	1.610	2.11	1.88	0.10
	FOMM [13]	1.085	0.424	12.28	1.626	3.29	2.92	0.18
	DPE [23]	1.577	0.528	14.45	1.581	1.62	1.63	0.05
	POCE-FE	0.924	**0.338**	10.73	1.652	3.57	3.70	**0.36**
	POCE-AU	**0.897**	0.366	**10.35**	**1.663**	**3.74**	**3.98**	0.31

it just requires unpaired images in training. Fig. 6 illustrates the editing of a few sample images from MultiPIE [44] and RaFD [45]. Specifically, we first use OpenFace [47] to extract the AUs intensities of source and reference images to derive relative AUs, then feed the source images as well as the relative AUs to POCE to produce the editing results. It can be seen that POCE can edit expressions and poses simultaneously without changing the facial identity, and even successfully generate extreme expressions as shown in the third column over MultiPIE [44]. More importantly, POCE edits expressions and poses in a disentangled manner, which translates to great editing flexibility by allowing to edit expressions only, poses only, or both of them with expressions and poses from different reference images. The superior editing usability and flexibility are largely attributed to the proposed disentangling approach within the UV space.

Facial Expression Editing via Feature Transfer: Beyond expression editing over relative AUs, POCE allows direct expression transfer (from reference UV to source UV) via feature transfer with given masks. Fig. 7 illustrates the editing of a few samples from MultiPIE [44] and RaFD [45]. Specifically, the trained UV completion network first generates the completed source and reference UV maps, which are then fed to the UV editing network (together with masks that indicate the regions to be edited) to produce the edited UV. Note that the mask could be in any shape, which allows to flexibly transfer expression information around interested facial regions from reference UV to source UV without any expression annotations. The superior editing flexibility is largely attributed to our proposed modulation-based generator.

Qualitative Evaluation: We first compare POCE with the state-of-the-art pose-fixed expression editing methods GANimation [5], MSF [7] and DPE [23] in Fig. 8(a), where non-facial regions are masked for better comparisons. POCE, GANimation [5] and MSF [7] are trained with continuous AUs intensities extracted by OpenFace [47], while DPE [23] is trained to learn disentangled pose and expression representations from video data. POCE-FE and POCE-AU indicate the editing results generated via feature transfer and conditioning on relative AUs, respectively. As Fig. 8(a) shows, GANimation [5], MSF [7] and DPE [23] tend to generate blurs and artifacts and even corrupted facial regions around mouths. Furthermore, their generated images show inconsistent expression intensity with the reference images (e.g., samples in column 1, 2 and 5). POCE can instead generate more realistic expressions with much less blurs and better consistency with

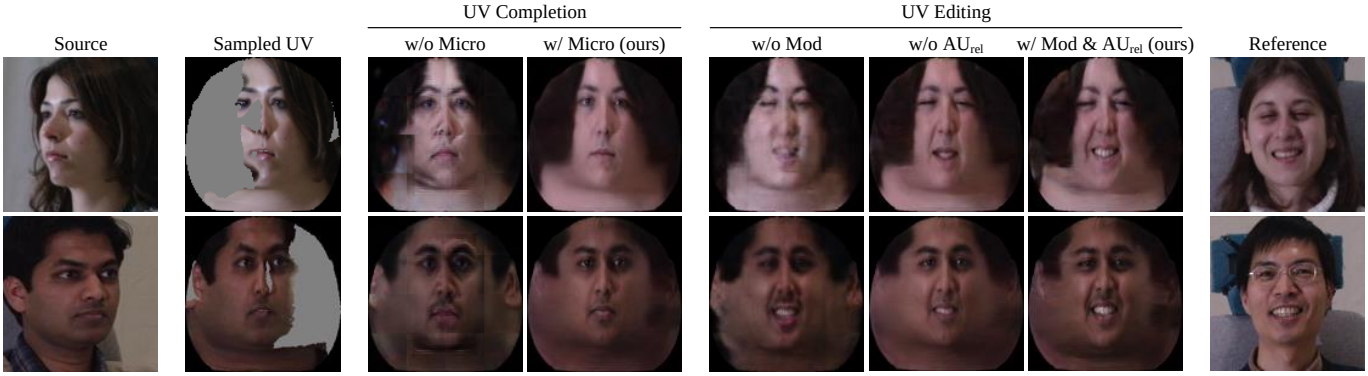


Fig. 9. Qualitative ablation study of POCE on MultiPIE: The proposed micro patch design helps generate better details and less artifacts as compared with *w/o Micro* in UV completion. In addition, including the modulation block (*Mod*) and relative AUs (AU_{rel}) helps generate more realistic facial details and more consistent expression intensity in UV editing.

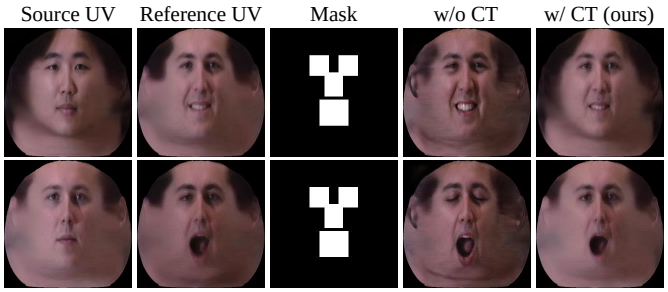


Fig. 10. Qualitative ablation study of POCE's expression editing via feature transfer on MultiPIE: Without using the constant tensor (CT), the produced UV maps are degraded by inconsistent expression intensity and altered person identity after feature transfer.

TABLE III
QUANTITATIVE ABLATION STUDY OF POCE ON MULTIPIE.

Models	IED ↓	EXD ↓	FID ↓	IS ↑
w/o Micro	0.872	0.535	25.19	1.581
w/o Mod	0.866	0.547	26.34	1.560
w/o AU$_{rel}$	0.828	0.513	19.78	1.631
w/o CT	0.839	0.506	20.12	1.625
POCE	**0.820**	**0.492**	**19.46**	**1.640**

the reference images. The better editing is largely attributed to our designed modulation-based generator that helps generate better facial details and edit expressions more precisely. In addition, GANimation [5] and MSF [7] cannot generate new poses in editing, while POCE can edit expressions and poses simultaneously as shown in Fig. 6.

We then compare POCE with FSTH [2], FOMM [13] and DPE [23] for pose-controllable expression editing in Fig. 8(b), where non-facial regions are masked for better comparisons. FSTH [2] and FOMM [13] are trained with paired data, and they generate edited faces conditioned on facial landmarks and the predicted optical flows, respectively. As Fig. 8(b) shows, FSTH [2] and FOMM [13] tend to generate degraded facial details and inconsistent expressions with reference images, and they fail to preserve facial identity in some editing (e.g., samples in column 2 and 5). One possible reason is landmarks and the predicted flows cannot capture fine-grained expression details and inevitably encode certain identity information of the reference face. DPE [23] struggles to generate realistic editing when the source and reference images have very different pose (e.g., samples in column 3, 5 and 6). POCE can instead generate more realistic expressions with less blurs and it preserves facial identity better as well. In addition, FSTH [2] and FOMM [13] edit expressions and poses in an entangled manner, and DPE [23] requires a reference image with desired pose and expression as guidance for editing. On the other hand, POCE allows to freely control pose and expressions using interpretable visual signals as illustrated in Fig. 6, which achieves much better editing flexibility.

Quantitative Evaluation: Table I shows quantitative comparisons of pose-fixed expression editing over MultiPIE [44] and RaFD [45]. We can see POCE consistently performs the best over all metrics. For identity preservation, POCE-AU achieved the best IED compared with the state-of-the-art methods. For expression editing accuracy, POCE-FE obtained the best EXD as it directly blends the expression-related features from reference to source UV, leading to more accurate expression editing. For perceptual quality, POCE-FE and POCE-AU consistently outperform other competitive methods on MultiPIE [44] and RaFD [45], with an improvement of FID by 2.21 and 0.68 as well as IS by 2.63% and 2.72%. Further, the POCE edited images obtained clearly higher scores in the user studies, which demonstrates the superiority of POCE in generating more realistic editing and more consistent expressions with respect to the reference images.

We also compare POCE with FSTH [2], FOMM [13] and DPE [23] over MultiPIE [44] and RaFD [45] with the same metrics. As Table II shows, POCE clearly outperforms the competing methods under different metrics, which suggests POCE can achieve more realistic synthesis in pose-controllable expression editing.

C. Discussion

Ablation Study: We first study how our proposed micro patch generation, modulation block and relative AUs contribute

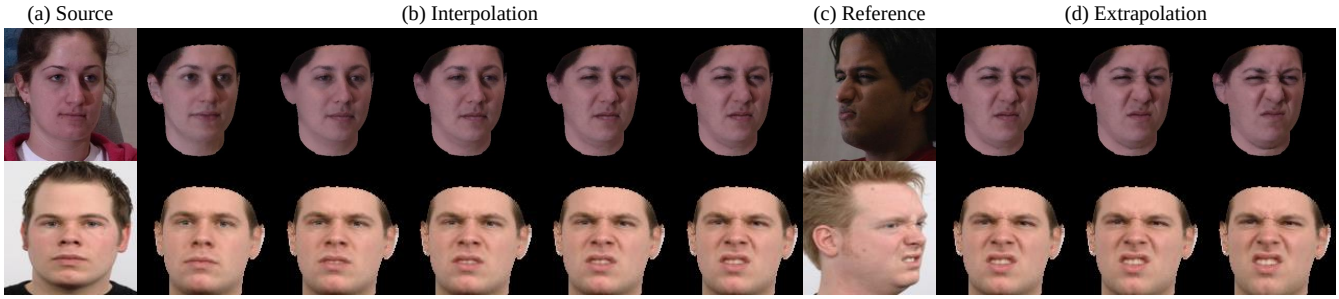


Fig. 11. Continuous expression editing by POCE: Given source images in (a) and reference images in (c), POCE can edit expressions by either interpolation or extrapolation over the relative AUs as shown in (b) and (d).

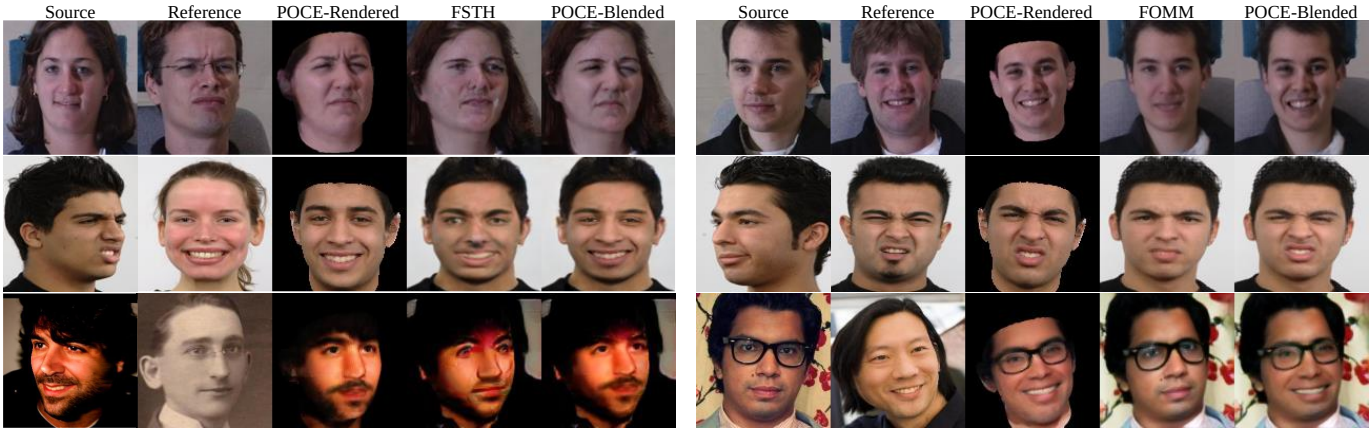


Fig. 12. Face blending by the proposed POCE over MultiPIE (first row), RaFD (second row) and wild images (last row): The POCE-rendered images can be seamlessly blended with FSTH [2] and FOMM [13] outputs to generate realistic images.

to the UV completion and UV editing. As Fig. 9 shows, including the proposed micro patch generation improves the UV completion in both smoothness and realism as compared with direct UV completion (w/o Micro) which produces inconsistent facial texture. In addition, the edited UV maps suffer from blurred facial details if we replace the modulation block with transposed convolutional layer (w/o Mod), and they tend to contain inconsistent expression intensity with the reference images (e.g., the mouth region) without using relative AUs (w/o AU_{rel}). Including the modulation block and relative AUs clearly produces sharper details and better expression consistency with the reference images.

We next study the contribution of the proposed constant tensor (CT) in expression editing via feature transfer. As described in Section III-C3, the constant tensor encodes coarse geometry prior that is shared among all UV maps. The generator takes the same constant tensor as initial input, gradually injects detailed texture information into it conditioning on the modulation parameters, and finally produces edited UV with desired expressions. In this process, the constant tensor serves as an anchor that helps the generator build spatial connection between the modulated parameters and the output UV map, which enables expression editing via feature transfer. The generator fails to learn such connection if we replace the constant tensor with randomly sampled Gaussian noise in network training (w/o CT), leading to inconsistent expression intensity and altered identity information in the produced UV maps as illustrated in Fig. 10.

We also conduct quantitative experiments to evaluate the contributions of each component. Table III shows the experimental results. The quantitative experimental results further verify the effectiveness of our proposed techniques.

Continuous Expression Editing: POCE can be easily adapted to generate continuous facial expressions. Given relative AUs between the source and the reference images, intermediate relative AUs of different stages can be simply derived by linear interpolation. Continuous facial expressions can thus be generated from the source images and the interpolated AUs by POCE as illustrated in Fig. 11.

Face Blending: As facial UV maps are sampled from ear-to-ear facial region, the POCE rendered images do not capture hair and image background. This limitation is shared among most existing 3D face editing methods [27], [42], [53]–[55]. We introduce face blending to mitigate this problem. Specifically, we first generate the target head pose with existing pose-controllable expression editing model, which usually contains corrupted facial features and inconsistent expression with the reference image, and then blend it with the POCE-rendered expression via Poisson editing [56]. In Fig. 12, we seamlessly blend the POCE-rendered faces to the outputs of FSTH [2] and FOMM [13] to generate realistic editing results.

Expression Editing on Wild Images: To evaluate the gener-

alization ability of our proposed method beyond lab settings, we further train our model on the 300W-LP dataset [39], which consists of 61,225 face images of various poses that are captured in both indoor and outdoor environments. As shown in the last row of Fig. 12, POCE successfully edits facial expressions and head poses successfully while maintaining identity information and personal attributes well (e.g., mustache and eyeglasses), demonstrating the robustness of POCE in handling diverse and uncontrolled images.

Ethical Considerations: With the convenience of generating pose-controllable expression editing faces from unpaired images, POCE could be misused by immoralists to spread misinformation. To avoid improper uses, we will include watermark to generated faces to indicate that they are synthetic.

V. Conclusion

This paper presents POCE for pose-controllable expression editing with just unpaired training data. Our method converts face images into UV maps and disentangles expression editing and head pose generation elegantly. We propose self-supervised UV completion and weakly-supervised UV editing to complete the missing facial textures on the sampled UV maps and edit expression information on the completed UV, respectively. Extensive experiments show that POCE can generate realistic facial expressions and head poses simultaneously. We expect POCE will inspire new insights and attract more interests for better expression editing in the near future.

VI. Acknowledgments

This work is funded by the Ministry of Education, Singapore, under the Tier-1 Project RG94/20 and the Tier-2 Project MOE-T2EP20220-0003.

References

[1] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012.

[2] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9459–9468.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[4] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

[5] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 818–833.

[6] R. Wu, G. Zhang, S. Lu, and T. Chen, "Cascade ef-gan: Progressive facial expression editing with local focuses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5021–5030.

[7] J. Ling, H. Xue, L. Song, S. Yang, R. Xie, and X. Gu, "Toward fine-grained facial expression manipulation," in *European Conference on Computer Vision*. Springer, 2020, pp. 37–53.

[8] V. Blanz, T. Vetter *et al.*, "A morphable model for the synthesis of 3d faces." in *Siggraph*, vol. 99, no. 1999, 1999, pp. 187–194.

[9] J. Geng, T. Shao, Y. Zheng, Y. Weng, and K. Zhou, "Warp-guided gans for single-photo facial animation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–12, 2018.

[10] J. Zhang, X. Zeng, M. Wang, Y. Pan, L. Liu, Y. Liu, Y. Ding, and C. Fan, "Freenet: Multi-identity face reenactment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5326–5335.

[11] E. Burkov, I. Pasechnik, A. Grigorev, and V. Lempitsky, "Neural head reenactment with latent pose descriptors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 786–13 795.

[12] S. Ha, M. Kersner, B. Kim, S. Seo, and D. Kim, "Marionette: Few-shot face reenactment preserving identity of unseen targets," *arXiv preprint arXiv:1911.08139*, 2019.

[13] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[14] D. Vlasic, M. Brand, H. Pfister, and J. Popovic, "Face transfer with multilinear models," in *ACM SIGGRAPH 2006 Courses*, 2006, pp. 24–es.

[15] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.

[16] Z. Geng, C. Cao, and S. Tulyakov, "3d guided fine-grained face manipulation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9821–9830.

[17] R. Wu and S. Lu, "Leed: Label-free expression editing via disentanglement," *arXiv preprint arXiv:2007.08971*, 2020.

[18] W. Wu, Y. Zhang, C. Li, C. Qian, and C. Change Loy, "Reenactgan: Learning to reenact faces via boundary transfer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 603–619.

[19] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.

[20] O. Wiles, A. Sophia Koepke, and A. Zisserman, "X2face: A network for controlling face generation using images, audio, and pose codes," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 670–686.

[21] C. Fu, Y. Hu, X. Wu, G. Wang, Q. Zhang, and R. He, "High fidelity face manipulation with extreme pose and expression," *arXiv preprint arXiv:1903.12003*, 2019.

[22] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4176–4186.

[23] Y. Pang, Y. Zhang, W. Quan, Y. Fan, X. Cun, Y. Shan, and D.-m. Yan, "Dpe: Disentanglement of pose and expression for general video portrait editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 427–436.

[24] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[25] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6721–6729.

[26] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4471–4480.

[27] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou, "Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7093–7102.

[28] B. Gecer, J. Deng, and S. Zafeiriou, "Ostec: One-shot texture completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7628–7638.

[29] C. H. Lin, C.-C. Chang, Y.-S. Chen, D.-C. Juan, W. Wei, and H.-T. Chen, "Coco-gan: Generation by parts via conditional coordinating," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4512–4521.

[30] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[31] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, 1978.

[32] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, "Stgan: A unified selective transfer network for arbitrary image attribute editing,"

in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3673–3682.

[33] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346.

[34] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[35] H. Kim, Y. Choi, J. Kim, S. Yoo, and Y. Uh, "Exploiting spatial dimensions of latent in gan for real-time image editing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[36] W. Peebles, J.-Y. Zhu, R. Zhang, A. Torralba, A. A. Efros, and E. Shechtman, "Gan-supervised dense visual alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 470–13 481.

[37] R. R. Althoff and N. J. Cohen, "Eye-movement-based memory effect: a reprocessing effect in face perception." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 25, no. 4, p. 997, 1999.

[38] J. H.-w. Hsiao and G. Cottrell, "Two fixations suffice in face recognition," *Psychological science*, vol. 19, no. 10, pp. 998–1006, 2008.

[39] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146–155.

[40] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 534–551.

[41] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1021–1030.

[42] H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang, "Rotate-and-render: Unsupervised photorealistic face rotation from single-view images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5911–5920.

[43] H. Kato, Y. Ushiku, and T. Harada, "Neural 3d mesh renderer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3907–3916.

[44] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and vision computing*, vol. 28, no. 5, pp. 807–813, 2010.

[45] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[46] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: adaptive curriculum learning loss for deep face recognition," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5901–5910.

[47] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[48] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[49] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *arXiv preprint arXiv:1606.03498*, 2016.

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[51] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[53] K. Nagano, J. Seo, J. Xing, L. Wei, Z. Li, S. Saito, A. Agarwal, J. Fursund, and H. Li, "pagan: real-time avatars using dynamic textures," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–12, 2018.

[54] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1155–1164.

[55] A. Lattas, S. Moschoglou, B. Gecer, S. Ploumpis, V. Triantafyllou, A. Ghosh, and S. Zafeiriou, "Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild"," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 760–769.

[56] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *ACM SIGGRAPH 2003 Papers*, 2003, pp. 313–318.