Explicit Facial Expression Transfer via Fine-Grained Semantic Representations

Zhiwen Shao\(^1\), Hengliang Zhu\(^1\), Junshu Tang\(^1\), Xuequan Lu\(^2\), Lizhuang Ma\(^1\)
\(^1\) Department of Computer Science and Engineering, Shanghai Jiao Tong University
\(^2\) School of Information Technology, Deakin University
{shaozhiwen, hengliang_zhu, tangjs}@sjtu.edu.cn, xuequan.lu@deakin.edu.au, ma-lz@cs.sjtu.edu.cn

Abstract

Facial expression transfer between two unpaired images is a challenging problem, as fine-grained expressions are typically tangled with other facial attributes such as identity and pose. Most existing methods treat expression transfer as an application of expression manipulation, and use predicted facial expressions, landmarks or action units (AUs) of a source image to guide the expression edit of a target image. However, the prediction of expressions, landmarks and especially AUs may be inaccurate, which limits the accuracy of transferring fine-grained expressions. Instead of using an intermediate estimated guidance, we propose to explicitly transfer expressions by directly mapping two unpaired images to two synthesized images with swapped expressions. Since each AU semantically describes local expression details, we can synthesize new images with preserved identities and swapped expressions by combining AU-free features with swapped AU-related features. To disentangle the images into AU-related features and AU-free features, we propose a novel adversarial training method which can solve the adversarial learning of multi-class classification problems. Moreover, to obtain reliable expression transfer results of the unpaired input, we introduce a swap consistency loss to make the synthesized images and self-reconstructed images indistinguishable. Extensive experiments on RaFD, MMI and CFD datasets show that our approach can generate photo-realistic expression transfer results between unpaired images with different expression appearances including genders, ages, races and poses.

Introduction

Facial expression transfer aims at transferring the expression from a source image to a target image, such that the transformed target image has the source expression while preserving other attributes like identity, pose and texture. It has recently gained remarkable attentions in the computer vision community (Ding, Sricharan, and Chellappa 2018; Song et al. 2018; Qiao et al. 2018; Pumarola et al. 2018). However, in literature preserving fine-grained details for expression transfer has remained a challenging problem, since fine-grained expressions are typically tangled with other facial attributes.

Recently, several facial expression manipulation methods have been proposed and can also be applied to expression transfer. Choi et al. (Choi et al. 2018) treated each discrete expression as a domain respectively, while Ding et al. (Ding, Sricharan, and Chellappa 2018) modeled the expression intensities to generate a wider range of expressions. An expression category only describes overall facial emotion, which has a limited capacity to capture fine details. Song et al. (Song et al. 2018) and Qiao et al. (Qiao et al. 2018) exploited facial landmarks to more finely guide the expression synthesis. However, transforming the landmarks from a source image to adapt to a target image with significantly different facial shape is difficult and may cause artifacts in the synthesized image.

Considering each facial action unit (AU) (Ekman and Rosenberg 1997) represents local muscle actions and can semantically describe fine-grained expression details, Pumarola et al. (Pumarola et al. 2018) took AUs with continuously varied intensities as a guidance to synthesize expressions. Nevertheless, only global AU features are learned, which limits the performance of editing local expressions. These methods all treat expression transfer as an application of expression manipulation, and require predicted expressions, landmarks or AUs of the source image to guide the expression edit of the target image. This offline prediction process is unnecessary and may degrade the performance of expression transfer due to the inaccurate predictions.

To tackle the above limitations, we propose to explicitly transfer fine-grained expressions by directly mapping two unpaired input images to two synthesized images with swapped expressions. The AU intensity estimation process is integrated into our framework by supervising the learning of AU-related features and the inheritance of AU information from the source image. Figure 1 shows the overview of our framework. In particular, the input images are first disentangled into two semantic representations (AU-related and AU-free features) by a novel adversarial training method. Compared to the conventional adversarial loss (Goodfellow et al. 2014) for binary classification problems only, our proposed adversarial training method can solve the adversarial learning of multi-class classification problems.

To capture fine expression details in each local region, we adopt an independent branch to extract a related local feature for each AU respectively and then combine these
denotes the disentanglement and generation parts, which consists of $DG$, flow mapping (Yang et al. 2011) and image techniques including 2D or 3D image warping (Garrido et al. 2014), flow mapping manipulation methods resorting to computer graphics techniques. 

Facial Expression Manipulation. There are many facial expression manipulation methods resorting to computer graphics techniques including 2D or 3D image warping (Garrido et al. 2014), flow mapping (Yang et al. 2011) and image rendering (Yang et al. 2012). Although these types of approaches can often generate realistic images with high resolution, the elaborated yet complex processes cause expensive computations. Recently, some works exploited the prevailing generative adversarial networks (GANs) (Goodfellow et al. 2014) to edit facial attributes including expressions.

Choi et al. (Choi et al. 2018) proposed a StarGAN method that can perform image-to-image translation for multiple domains using only a single model. It shows a superiority in facial attribute transfer and expression synthesis, but only eight emotional expressions were synthesized. Another work (Ding, Sricharan, and Chellappa 2018) designed an Expression Generative Adversarial Network (ExprGAN) for expression edit with controllable expression intensities. To control finer details, Pumarola et al. (Pumarola et al. 2018) utilized AUs as the guidance to synthesize expressions in a continuous domain. This approach allows controlling the intensity of each AU and combining several of them to synthesize an expression. However, only global AU features are learned for expression synthesis, which limits the performance of editing local details. Considering the geometry characteristics of expressions, Song et al. (Song et al. 2018) and Qiao et al. (Qiao et al. 2018) proposed Geometry-Guided GANs to generate expressions with the geometry formed by facial landmarks. Nevertheless, transforming source-image landmarks to match a target image with significantly different facial shape is difficult and usually causes artifacts in the generated image. These methods all require the predictions of expressions, AUs or landmarks and cannot estimate them automatically.

**Feature Disentanglement.** Similar to previous works (Tran, Yin, and Liu 2017; Lee et al. 2018; Shu et al. 2018), our method also uses feature disentanglement to factorize an image into different representations by GANs. Each disentangled representation is distinct and can be specialized for a

![Figure 1: The overview of our EET framework. (a) During training, $E$ extracts facial features from two unpaired input images. $DG$ denotes the disentanglement and generation parts, which consists of $E_r$, $E_f$ and $G$, as illustrated in (b). $E_r$ and $E_f$ disentangle the facial features into AU-related features ($\mathbf{f}_r^a$, $\mathbf{f}_r^b$) and AU-free features ($\mathbf{f}_f^a$, $\mathbf{f}_f^b$), respectively. $G$ combines the AU-free features with the swapped AU-related features to generate two new images $\mathbf{I}^a$ and $\mathbf{I}^b$. A further disentanglement-swap-generation process is conducted to cross-cyclically reconstruct the input images. At test time, the two input images only go through the components in the rectangular box.](image)

Features as the AU-related feature. After the feature disentanglement, the AU-related features of the two images are swapped and combined with the AU-free features to generate two new images with swapped expressions. To enable the reliability of expression transfer for unpaired input, we introduce a swap consistency loss to make the generated images and self-reconstructed images indistinguishable. Another disentanglement-swap-generation process is further applied to the generated images to complete the crossed cycle. At test time, taking two unpaired images as input, our method automatically outputs two synthesized images with swapped expressions.

We refer to our framework, explicit expression transfer, as **EET**. The main contributions of this paper are summarized as follows. First, we propose a novel explicit facial expression transfer framework to transfer fine-grained expressions between two unpaired images. Second, we propose an adversarial training method to disentangle the AU-related feature and the AU-free feature, which can be applied to the adversarial learning of multi-class classification problems. Third, we introduce a swap consistency loss to ensure the reliability of expression transfer for unpaired input. Finally, extensive experiments on benchmark datasets demonstrate that our approach can generate photo-realistic expression transfer results between unpaired images with various expression appearance differences.

**Related Work**

We review previous techniques that are closely related to our work, in terms of facial expression manipulation and feature disentanglement.
certain task. Tran et al. (Tran, Yin, and Liu 2017) proposed a disentangled representation learning GAN for pose-invariant face recognition. Lee et al. (Lee et al. 2018) utilized the disentangled features to produce diverse and realistic images, and also employed a cyclic structure (Zhu et al. 2017) to deal with unpaired training data. Shu et al. (Shu et al. 2018) proposed a generative model which disentangles shape from appearance in an unsupervised manner. In this method, shape is represented as a deformation, and appearance is modeled in a deformation-invariant way. These methods enforce the disentangled feature to be close to a prior distribution, or exploit an opposite feature containing specific information to implicitly encourage the disentangled feature to discard the information. As a result, they have limited applicability.

In contrast, our approach can disentangle representations for any multi-class classification problems by a novel adversarial training method.

Explicit Facial Expression Transfer

Overview

Given two unpaired input images \((I^a, I^b)\), our main goal is to generate two new images \((\hat{I}^a, \hat{I}^b)\) with swapped facial expressions while preserving other original attributes like identity, pose and texture. AU intensity labels \((u^a, u^b)\) and identity labels \((d^a, d^b)\) of the two input images are provided, and pose labels \((p^a, p^b)\) are also available if the training set contains images with different poses. Taking \(I^a\) as an example, \(u^a = (u^a_1, \ldots, u^a_m)\) denotes the intensities of \(m\) AUs, where \(u^a_i \in [0, 1], i = 1, \ldots, m\) represents the maximum intensity level. \(d^a \in \{1, \ldots, n\}\) and \(p^a \in \{1, \ldots, v\}\), where \(n\) and \(v\) are the numbers of identity classes and pose classes, respectively.

Figure 1 illustrates the overall architecture of our EET framework. During training, for two unpaired input images \((I^a, I^b)\), we first utilize a feature encoder \(E\) to extract their facial features \((f^a, f^b)\) which contain rich information such as expression, identity, pose and texture. As shown in Figure 1(b), \(D\) consists of a feature disentanglement process with an AU-related encoder \(E_r\) and an AU-free encoder \(E_f\), and an image generation process with a generator \(G\). Specifically, \(E_r\) and \(E_f\) disentangle \((f^a, f^b)\) into AU-related features \((\hat{f}^a, \hat{f}^b)\) and AU-free features \((f^a_\ell, f^b_\ell)\), respectively. \(G\) further combines the AU-free features with the swapped AU-related features to generate new images \((\hat{I}^a, \hat{I}^b)\). Then, another disentanglement-swap-generation process is applied to generate the cross-cyclically reconstructed images \((\hat{I}^a, \hat{I}^b)\). The key to our proposed explicit expression transfer lies in combining AU-free features with swapped AU-related features to generate photo-realistic images.

Expression Transfer

Explicit expression transfer requires the disentanglement of two semantic representations: AU-related feature and AU-free feature. To remove AU information for the AU-free feature, an alternative solution is to use the typical GAN with a two-player minimax game (Goodfellow et al. 2014) to adversarially train the AU-free encoder \(E_f\) and an AU discrimi-

\begin{equation}
\mathcal{L}_{u}^{(D)}(I^a) = \mathbb{E}_{D}(\frac{1}{m(l+1)} \sum_{i=1}^{m} \sum_{q=0}^{l} [\mathbb{I}_{[q \neq u^a_i]} ||D_{(u^a_i,q)}(f^a_i)||^2_2 + \mathbb{I}_{[q = u^a_i]} ||D_{(u^a_i,q)}(f^a_i) - 1||^2_2]).
\end{equation}

\begin{equation}
\mathcal{L}_{u}^{(E)}(I^a) = \mathbb{E}_{f}(\frac{1}{m(l+1)} \sum_{i=1}^{m} \sum_{q=0}^{l} [||D_{(u^a_i,q)}(f^a_i)||^2_2 - \frac{1}{l+1} ||^2_2]),
\end{equation}

where \(D_{u} \) so that \(D_{u} \) cannot discriminate AU attribute from the output of \(E_{f}\). Since AU intensity estimation is a multi-label regression problem, we regard it as a multi-label multi-class classification problem by discretizing the AU intensity labels:

\begin{equation}
u^a_i = \lfloor u^a_i \rfloor \in \{0, \ldots, l\},
\end{equation}

where \(\lfloor \cdot \rfloor\) denotes the operation of rounding a number to the nearest integer. However, the two-player minimax game is designed for binary classification problems, and cannot work for multi-class classification problems.

To solve this issue, we propose an adversarial training method with two stages, as illustrated in Figure 2. In the first training stage, our goal is to jointly train \(E_{f}, E_{r}, D_{u}\) and \(D_{d}\) so that \(D_{u}\) and \(D_{d}\) have certain abilities of classifying AU intensities and identities, respectively. Given \(I^a = \hat{I}^a\), we first obtain \(f^a_\ell = E_f(I^a)\) and \(f^a_r = E_r(I^a)\), which are further input to \(D_{u}\) and \(D_{d}\), respectively. \(D_{u}\) outputs an \(m(l+1)\)-dimensional vector with an AU intensity discrimination loss \(\mathcal{L}_{u}^{(D)}\):

Figure 2: The procedure of our proposed adversarial training. \(D_{u}\) and \(D_{d}\) with dotted lines denote that their parameters are fixed in the second stage.
the identity will not be transferred from $\mathbf{I}^a$ to $\mathbf{I}^b$. Similarly, we can define an identity discrimination loss $L_{ad}^{(D)}$ and an identity confusion loss $L_d^{(E)}$ which are used in the first and second training stages, respectively. The adversarial training between $E_r$ and $D_r$ encourages $f_r^a$ to be identity-free.

Moreover, to ensure $f_r^a$ contains AU-related information, an AU intensity estimation loss $L_u$ is applied to $E_r$. To capture fine-grained expression details in each local region, $E_r$ uses an independent branch to extract a related local feature for each AU separately. The top of each branch in $E_r$ is a convolutional layer followed by a global average pooling (Lin, Chen, and Yan 2014) layer and a one-dimensional fully-connected layer, in which the convolutional layer outputs $f_r^a$ and a sigmoid function is applied to the fully-connected layer to predict the normalized intensity $\hat{u}_r^a \in [0,1]$. Following the loss weighting strategy (Shao et al. 2018) to suppress the data imbalance issue, we introduce a weighted L2 loss for $L_u$:

$$L_u(\mathbf{I}^a) = \mathbb{E}_{\mathbf{I}^a} \left[ \frac{1}{m} \sum_{i=1}^{m} w_i \| u_i^a - \hat{u}_r^a \|_2^2 \right], \quad (4)$$

where $w_i = (1/o_i)/\sum_{i=1}^{m} (1/o_i)$ is the weight parameter. $o_i$ is the occurrence rate of the $i$-th AU in the training set, in which the AU intensities greater than $(l-1)/2$ are treated as occurrence and non-occurrence otherwise. By integrating local features of all the AUs, we can obtain a fine-grained semantic representation for $f_r^a$:

$$f_r^a = \frac{1}{m} \sum_{i=1}^{m} \odot f_r^a_i,$$  

where $\odot$ denotes element-wise sum.

As shown in Figure 1, two new images $\hat{\mathbf{I}}^a$ and $\hat{\mathbf{I}}^b$ with swapped expressions are generated through $G$: $\mathbf{I}^a = G(f_r^a, f_f^a)$ and $\mathbf{I}^b = G(f_r^b, f_f^b)$, in which the channels of $f_r^a$ and $f_r^b$ are concatenated as input to $G$. Since there are no ground-truth expression transfer results for two unpaired images, we introduce a swap consistency loss to ensure the reliability of generated images:

$$L_{sc}(\mathbf{I}^a) = \mathbb{E}_{\mathbf{I}^a} [\log D_{sc}(\hat{\mathbf{I}}^a)] + \mathbb{E}_{\mathbf{I}^a} [\log(1 - D_{sc}(\hat{\mathbf{I}}^a))], \quad (6)$$

where $L_{sc}$ encourages $\mathbf{I}^a$ and the self-reconstructed image $\hat{\mathbf{I}}^a = G(f_r^a, f_f^a)$ to be indistinguishable with a swap consistency discriminator $D_{sc}$. To constrain $\mathbf{I}^a$ to have the transferred expression from $\hat{\mathbf{I}}^b$ while preserving other attributes, we apply $L_u$ in Eq. (4) to $\mathbf{I}^a$ with label $\mathbf{I}^b$. Besides, the facial feature $\mathbf{f}^a = E(\mathbf{I}^a)$ of $\mathbf{I}^a$ is input to an attribute constraint module $C$ which outputs an $n$-dimensional vector and a $v$-dimensional vector. We formulate the attribute constraint loss as

$$L_c(\mathbf{I}^a) = -\mathbb{E}_{\mathbf{I}^a} \left[ \sum_{j=1}^{n} \mathbb{I}_{[j = d^a]} \log(\sigma(C^{d^a}(\mathbf{f}^a))) + \lambda_p \sum_{k=1}^{v} \mathbb{I}_{[k = p^a]} \log(\sigma(C^{p^a}(\mathbf{f}^a))) \right],$$  

where $\lambda_p$ is a hyper-parameter for the trade-off between the first term of identity classification and the second term of pose classification, $C^{d^a}(\cdot)$ and $C^{p^a}(\cdot)$ denote the $j$-th identity value and the $k$-th pose value respectively, and $\sigma(\cdot)$ denotes a softmax function.

**Full Objective**

To facilitate the disentanglement-swap-generation process, a reconstruction loss (Lee et al. 2018) is employed:

$$L_r(\mathbf{I}^a) = E_{\mathbf{I}^a} [\| \hat{\mathbf{I}}^a - \mathbf{I}^a \|_1] + \| \hat{\mathbf{I}}^a - \mathbf{I}^a \|_1, \quad (8)$$

where the first and second terms constrain the self-reconstruction and cross-cycle reconstruction, respectively. To make the synthesized images look real and indistinguishable from the original images, we impose an image adversarial loss with an image discriminator $D_r$:

$$L_{ad_r}(\mathbf{I}^a) = \mathbb{E}_{\mathbf{I}^a} [\log D_r(\mathbf{I}^a)] + \mathbb{E}_{\mathbf{I}^a} [\log(1 - D_r(\mathbf{I}^a))]. \quad (9)$$

For stable adversarial training, we use the least-squares loss (Mao et al. 2017) to train $L_{sc}$ and $L_{ad_r}$.

In our EET framework, the losses introduced above are applied for both two input images so that expressions can be mutually transferred. During the first training stage, we jointly train $E, C, E_f, E_r, D_u$ and $D_d$. The full objective function is formulated as

$$\min_{\{E_f, E_r, G\}} \max_{\{D_{sc}, D_{ad}\}} \mathcal{L}_{EET} = L_u + \lambda_c L_c + \lambda_{ad_u} L_{ad}^{(D)} + \lambda_{ad_d} L_{ad}^{(D)}, \quad (10)$$

where $L_c$ is applied for input images to train the attribute constraint module $C$. During the second training stage, we train the overall framework shown in Figure 2 by fixing the parameters of $E, C, D_u$ and $D_d$. The full objective function is formulated as

$$\min_{\{E_f, E_r, G\}} \max_{\{D_{sc}, D_{ad}\}} \mathcal{L}_{EET} = L_u + \lambda_c L_c + \lambda_r L_r + \lambda_{ad_u} L_{ad}^{(E)} + \lambda_{ad_d} L_{ad}^{(E)} + \lambda_{sc} L_{sc} + \lambda_{ad} L_{ad},$$  

where $L_u$ is imposed to $\mathbf{I}^a$ and $\mathbf{I}^a$ for supervising the learning of $f_r^a$ and the inheritance of expression from $\mathbf{I}^a$ respectively, and the hyper-parameters $\lambda_{\cdot}$ weigh the importance of each loss term. At inference time, the two input images are simply disentangled and swapped to synthesize new images with swapped expressions.

**Experiments**

**Datasets and Settings**

**Datasets.** We evaluate our framework on three facial expression benchmark datasets: RaFD (Langner et al. 2010), MMI (Pantic et al. 2005; Valstar and Pantic 2010) and CFD (Ma, Correll, and Wittenbrink 2015). RaFD and CFD do not contain AU labels, and MMI only annotates a few AUs for a small set of images. To enable the training and testing on these datasets, we employ a powerful AU recognition library OpenFace (Baltrusaitis, Mahmoud, and Robinson 2015) to annotate continuous intensities of 12 AUs (1, 2, 4, 5, 6, 9, 12, 15, 17, 20, 25 and 26) for each image of these datasets.
RaFD consists of 67 subjects with 8 facial expressions, 3 gaze directions and 5 camera angles, which exhibits diverse expression variations. We select the samples recorded by the camera angles of $45^\circ$, $90^\circ$ and $135^\circ$ to evaluate expression transfer between images with pose differences, with a total number of 4,824 images. These images are randomly partitioned into a training set with 4,320 images of 60 subjects and a test set with 504 images of 7 subjects.

MMI contains 2,390 videos and 493 images of 67 subjects. We remove the videos under undesirable situations like severe side views and very low quality, and obtain 192,131 images of 43 subjects. These images are randomly partitioned into a training set with 173,374 images of 39 subjects and a test set with 18,757 images of 4 subjects.

CFD includes 1,207 images of 597 subjects, in which 158 subjects are represented with neutral, happy, angry and fearful expressions, and the remaining subjects only have a neutral expression. We randomly divide these images into a training set with 1,081 images of 537 subjects and a test set with 126 images of 60 subjects.

Implementation Details. We utilize PyTorch to implement our EET framework. Our framework comprises of $E$, $C$, $E_f$, $E_r$, $G$, $D_u$, $D_d$, $D_{sc}$ and $D_g$. $G$ contains 6 up-sampling residual blocks (Zheng, Cham, and Cai 2019), and $D_{sc}$ and $D_g$ are based on the structure of PatchGAN (Isola et al. 2017). Other modules are mainly composed of stacked convolutional layers proposed by VGGNet (Simonyan and Zisserman 2015). To obtain stable adversarial training, we conduct Spectral Normalization (Miyato et al. 2018) on each convolutional layer and deconvolutional layer in $E_f$, $E_r$, $G$, $D_u$, $D_d$, $D_{sc}$ and $D_g$.

In our experiments, each image is cropped to the size of $256 \times 256$ and further randomly mirrored for data augmentation. The number of AUs $m$ and the maximum intensity level $l$ are 12 and 5, respectively. The hyper-parameters of different loss terms are empirically set as: $\lambda_c = 2$, $\lambda_p = 1$, $\lambda_{adv} = 40$, $\lambda_{adv} = 40$, $\lambda_{sc} = 0.2$, $\lambda_{adv} = 1.2$ and $\lambda_r = 40$. If the training set involves only one pose, $\lambda_p$ is set to 0.

We employ the Adam solver (Kingma and Ba 2015), and set $\beta_1 = 0.95$, $\beta_2 = 0.999$ and an initial learning rate of $10^{-4}$ during the first training stage, as well as set $\beta_1 = 0.5$, $\beta_2 = 0.9$ and an initial learning rate of $3 \times 10^{-5}$ during the second training stage. For each stage, the learning rate is kept unchanged during the first half of training epochs and linearly decayed at each epoch during the remaining half of training epochs. We use all the RaFD, MMI and CFD training images in the first stage, and then use each dataset to train individual models. We utilize 140, 4 and 350 epochs for RaFD, MMI and CFD respectively, in terms of both the first and second stages.

Evaluation Metrics. To quantitatively evaluate our method for expression transfer, we use Pearson’s correlation coefficient (PCC) and mean square error (MSE) to measure the correlation and difference between AU intensities of source images and generated target images, respectively. The average results of PCC and MSE over all AUs (Avg) are also shown. Besides, we evaluate identity preservation by conducting face verification between real images and generated images, with accuracy and true accept rate at 1% false accept rate (TAR@FAR=1%) reported.

Ablation Study

In this section, we evaluate the main loss terms in our framework. We design a baseline method B-Net using the same architecture and two training stages as EET but with only the losses $L_u$, $L_r$ and $L_{adv}$. Other variants of EET are built based on B-Net by further adding $L_c$, and all of $L_u^{(D)}$, $L_d^{(D)}$, $L_u^{(E)}$ and $L_d^{(E)}$, denoted as BC-Net, and BCA-Net, respec-
We also implement a variant named EET-Single by using a single branch for all the AUs in the structure of $E_{rv}$, in which the single branch outputs $f_a$ followed by a global average pooling layer and an $m$-dimensional fully-connected layer. Their results for two example pairs of RaFD images are illustrated in Figure 3.

We can see that B-Net transfers the entire face from $I^a$ to $I^b$ rather than only the expression, and BC-Net still fails to preserve the identity during expression transfer after applying the attribute constraint loss $L_c$. Although B-Net and BC-Net employ the commonly used $L_c$ to obtain good self-reconstructed results $I^a$ and $I^b$, there are many artifacts in the expression transfer results. When using our proposed adversarial training method to disentangle the AU-related feature and the AU-free feature, BCA-Net significantly improves the performance of expression transfer, but expressions are not completely swapped and a few artifacts appear around the mouth regions, especially for images with large pose differences. By employing the swap consistency loss $L_{sc}$ to make expression transfer and self-reconstruction indistinguishable, our EET generates photo-realistic expression transfer results while preserving original identities for unpaired images with different genders and poses.

When using a single branch for $E_{rv}$, we can observe that the generated $I^b$ by EET-Single of the first pair still looks a little happy rather than angry, and the closed lips looks unnatural. In addition, EET generates more realistic mouth regions in $I^b$ of the second pair than EET-Single. This demonstrates that a single branch for extracting global AU features has a limited capacity to capture fine-grained expression details. We also show the expression transfer results of our EET on example images of MMI and CFD in Figure 4. It can be seen that our method is capable of transferring fine details such as glances and local muscle actions. Although MMI images are a little blurry and may be partially occluded by eyeglasses, our approach can successfully transfer expressions and synthesize high-quality images. By observing the results on CFD, we can see that our EET is robust to unpaired images with different ages and races.

### Comparison with State-of-the-Art Methods

To validate our framework, we compare it against two state-of-the-art facial expression manipulation methods, StarGAN (Choi et al. 2018) and HomoInterpGAN (Chen et al. 2019). We implement these two methods by running their released code on RaFD dataset. Since the combination of AUs with certain intensities describes an expression, multiple AUs with continuously varied intensities can obtain an infinite number of expressions. Thus, StarGAN and HomoInterpGAN which are designed for discrete expressions cannot process continuously varied AUs. In our experiments, StarGAN and HomoInterpGAN are trained using the expression labels, in which HomoInterpGAN can directly transfer the expression from a source image to a target image and StarGAN generates a new target image by giving the expression label of a source image. Note that a few other works with code released such as ExprGAN (Ding, Sricharan, and Chellappa 2018) and GANimation (Pumarola et al. 2018) are not compared, because their code hardly generates valid results when training on RaFD.

### Qualitative Results

Figure 5 presents the expression transfer results of these methods between unpaired images with both identical poses and different poses on RaFD. It can be observed that the results of StarGAN for some pairs are a little blurry with shadows, especially for the lips. The expressions are partially transferred such as $I^b$ of the first pair in Figure 5(a) showing mixed expressions of fear and disgust. Although HomoInterpGAN generates images with higher quality than StarGAN, the identities of target images are significantly changed. For example, it seems that the gender of $I^b$ of the third pair in Figure 5(a) is changed from male to female. By contrast, our method EET is able to transfer expressions to generate realistic images while preserving original identity information. Moreover, EET has a stronger ability of transferring fine expression details than StarGAN and HomoInterpGAN such as lifting one side of the lip corners.
Figure 5: Comparison of expression transfer results using StarGAN, HomoInterpGAN and our EET for six pairs of RaFD images. (a) The expression pairs are (disgusted, fearful), (happy, neutral) and (disgusted, surprised), respectively. (b) The expression pairs are (contemptuous, neutral), (happy, angry) and (angry, surprised), respectively.

Table 2: Quantitative results (%) of identity preservation.

| Method        | Accuracy | TAR@FAR=1% |
|---------------|----------|------------|
| StarGAN       | 95.63    | 87.16      |
| HomoInterpGAN | 70.82    | 17.02      |
| EET           | 96.63    | 90.90      |
| Real          | 98.23    | 97.86      |

Conclusion

In this paper, we have proposed an EET framework to explicitly transfer fine-grained expressions by straightforwardly mapping the unpaired input to two synthesized images with swapped expressions. We have also proposed an adversarial training method to disentangle AU-related features and AU-free features, which can solve the adversarial learning of multi-class classification problems. Moreover, we have introduced a swap consistency loss to ensure the reliability of expression transfer. Extensive experiments show that our approach can generate photo-realistic expression transfer results while preserving identity information, even in the presence of challenging expression appearance differences.
Conference on Computer Vision and Pattern Recognition, 861–868. IEEE.

[Zheng, Cham, and Cai 2019] Zheng, C.; Cham, T.-J.; and Cai, J. 2019. Pluralistic image completion. In IEEE Conference on Computer Vision and Pattern Recognition, 1438–1447. IEEE.

[Zhu et al. 2017] Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In IEEE International Conference on Computer Vision, 2223–2232. IEEE.