All-In-One: Facial Expression Transfer, Editing and Recognition Using A Single Network

Kamran Ali, Charles E. Hughes
Synthetic Reality Lab, Department of Computer Science
University of Central Florida, Orlando, Florida
kamran@knights.ucf.edu, ceh@cs.ucf.edu

Abstract

In this paper, we present a unified architecture known as Transfer-Editing and Recognition Generative Adversarial Network (TER-GAN) which can be used: 1. to transfer facial expressions from one identity to another identity, known as Facial Expression Transfer (FET), 2. to transform the expression of a given image to a target expression, while preserving the identity of the image, known as Facial Expression Editing (FEE), and 3. to recognize the facial expression of a face image, known as Facial Expression Recognition (FER). In TER-GAN, we combine the capabilities of generative models to generate synthetic images, while learning important information about the input images during the reconstruction process. More specifically, two encoders are used in TER-GAN to encode identity and expression information from two input images, and a synthetic expression image is generated by the decoder part of TER-GAN. To improve the feature disentanglement and extraction process, we also introduce a novel expression consistency loss and an identity consistency loss which exploit extra expression and identity information from generated images. Experimental results show that the proposed method can be used for efficient facial expression transfer, facial expression editing and facial expression recognition. In order to evaluate the proposed technique and to compare our results with state-of-the-art methods, we have used the Oulu-CASIA dataset for our experiments.

1. Introduction

Facial Expression synthesis and manipulation is a challenging task because it requires the disentanglement of facial expression features from identity information. It has recently gained a great deal of attention from the computer vision research community due to the exciting research challenges it offers apart from its many applications, e.g facial animation, human-computer interactions, enter-
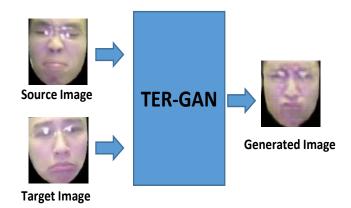

Figure 1. TER-GAN takes source image and target image as input and extracts expression information and identity information from each image respectively. These encoded information is then used to generate an expression image which contains the expression of source image while the identity of target image is preserved.

tainment and facial reenactment [32].

Many techniques have been developed for facial expression manipulation and editing. These techniques can be divided into two categories: graphic-based techniques [38], [21], [40] and generative methods [31], [26], [3], [6], [28], [22], [17]. In the first category, image warping is used to synthesize expression images by modeling the variations of a face during facial expressions. In the second category, deep generative models are used to generate synthesized expression images. In [6] an expression controller module is trained using a GAN-based architecture to generate expression images of various intensities. Similarly in [4], a unified GAN framework is used to transfer expressions from one image domain to another. Kamran et al. [1] concatenated a one-hot identity vector with the expression information extracted from the encoder to generate an ex-

pression image. Pumarola et al. [23] and Shao et al. [28] exploited Facial Action Units (AU), and the expression synthesis process is guided by the learned AU features. Similarly, in [30] and [25], facial landmarks are used to produce synthesized expression images.

Existing facial expression synthesis techniques have the capability to transform the expression of a given image; however, there are two main problems with these methods: 1. they require auxiliary information in the form of an expression code, facial landmarks and action unit information to synthesize an expression image and 2. many of these techniques fail to preserve the identity information of the given image, which is due to the fact that they fail to disentangle expression features from identity representation. Hence, during facial expression transfer process the identity information of the source image is usually leaked through the expression feature vector, which degrades the identity of generated images [36]. Similarly, in [30] and [25], it is very difficult to synthesize an expression image using the landmark information of source image with a different facial shape than the target face. To reduce the identity information leakage, in [1] and [6], the expression synthesis process is conditioned on expression and identity codes, rather than using the extracted features. But it is a well known fact that the identity and expression representations are too rich to be represented by one-hot vectors.

In order to overcome the above problems, we propose a Transfer-Editing and Recognition Generative Adversarial Network (TER-GAN) to automatically and explicitly extract a disentangled expression representation from a source image and disentangled identity features from a target image in order to synthesize a photo-realistic expression image without requiring any auxiliary information such as expression or identity code, facial landmarks or action units, while preserving the identity of the target image. The overall framework of our proposed technique is shown in Figure 1. TER-GAN has two main objectives: 1. to automatically extract disentangled expression features and identity features from a source image and a target image respectively, and 2. to synthesize a photo-realistic expression image containing the expression of the source image while preserving the identity of the target image.

To achieve these objectives, we employ a Generative Adversarial Network (GAN) with an encoder-decoder based generator G. As opposed to previous generative model based expression synthesis and manipulation architectures [6], [17], [35], we, instead of using just one encoder, employ two encoders G_{es} and G_{et} in our generator G. TER-GAN takes two images as input, source image x_s and target image x_t. Encoder G_{es} is aimed to encode expression representation $f(e)$ from source image x_s and encoder G_{et} is used to extract identity features $f(i)$ from target image x_t. The expression representation $f(e)$ is then concate-

nated with the identity feature $f(i)$: $f(x) = f(e) + f(i)$, and the concatenated feature vector $f(x)$ is then fed to the decoder G_{de} to synthesize an expression image $\bar{x}$, which contains the expression of source image x_s while preserving the identity of target image x_t. In order to further improve the quality of extracted expression and identity features, we make use of synthetic expression images along with real images, and introduce two adversarial losses at the output of each encoder: an adversarial expression consistency loss and an adversarial identity consistency loss. Our experimental results show that these two consistency losses help in extracting effective expression and identity features through which we can generate synthetic images that preserve the identity of the target image. Moreover, to generate more realistic synthesized expression images we use a multi-class classifier as our discriminator, D. The main contributions of our paper are as follows:

- We present a novel unified architecture, Transfer-Editing and Recognition Generative Adversarial Network (TER-GAN) that can be used efficiently for three purposes: facial expression transfer, facial expression editing and facial expression recognition, without requiring any explicit expression or identity code or any other auxiliary information such as facial landmarks or action units to guide the synthesis process, while preserving the identity information.

- Apart from the encoder-decoder architecture of TER-GAN, our adversarial expression and identity consistency losses also ensure that the expression and identity features are disentangled, and this disentanglement of features helps in synthesizing expression images that preserve the identity information of the target image.

- In order to deal with small expression datasets, TER-GAN learns to extract expression and identity representations using the information contained in synthesized images as well.

- We show that the disentangled expression embedding learned by TER-GAN can be effectively used for facial expression recognition.

2. Related Work

In this section we first review previous facial expression synthesis and manipulation techniques, followed by discussing conventional feature disentanglement techniques.

2.1. Facial Expression Manipulation

There are many facial expression manipulation techniques proposed in the literature, some of which combine
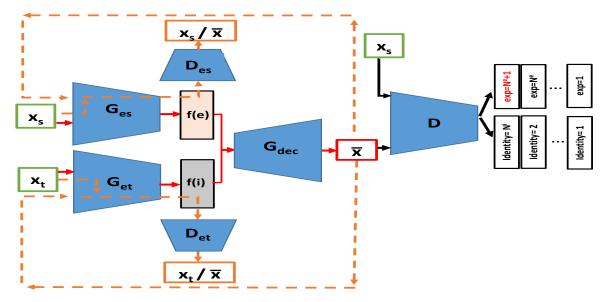
Figure 2. The over-all architecture of our TER-GAN (best viewed in color)

computer graphics methods such as 2D or 3D image warping [9], flow mapping [38] and image rendering [37] with computer vision algorithms to generate synthesized expression images. Although these techniques are able to produce photo-realistic high resolution synthesized images, the main problem with these approaches is that they suffer from high computation cost. In contrast to conventional facial expression synthesis techniques, recently proposed methods are mostly based on conditional Generative Adversarial Networks (cGANs).

Some earlier techniques such as [16], [14], and [42] used deterministic target expressions as one-hot vectors and generated synthesized images conditioned on discrete facial expressions. While the generated images are of better quality, these techniques are only able to generate discrete expressions. To overcome this problem, Ding et al. [6] proposed an Expressive GAN (ExprGAN) to synthesize an expression image conditioned on a real-valued vector that contains more complex information such as intensity variation. Similarly in [30] and [24], the image synthesis process is conditioned on geometry information in the form of facial landmarks. Choi et al. [5] proposed the StarGAN method to employ domain information and generate an image into a corresponding domain. In [5], Royer et al. presents XGAN to translate attributes from one domain to another by using an adversarial classifier on top of encoded layers, but their architecture is too simple to generate photo-realistic expression images. In another work, Pumarola et al. [42] used AUs as a conditional label to synthesize an expression image. All of these proposed techniques require explicit expression, AU and landmark information to guide the expression synthesis process.

2.2. Feature Disentanglement

Many techniques have been developed in the past to disentangle an image into its different representations. These techniques are based on the idea of learning by reconstruction, and therefore, often involve encoder-decoder structure coupled with GANs. For instance Tran et al. [33] proposed a disentangled representation learning GAN for pose-invariant face recognition in which the face identity representation is disentangled from the pose information by explicitly providing the pose information to the generator. Similarly, Lee et al. [15] generated photo-realistic images from various domains by disentangling the factors of variations in an input image. Shu et al. [29] proposed an unsupervised generative model to disentangle shape from appearance information. All of these methods learn the disentangled representation by explicitly providing information about other (uninterested) factors of variation that constitute an image. In contrast, our method automatically extracts the factors of variations from input images using two augmenting feature learning techniques, i.e learning by reconstruction employing an encoder-decoder based GAN set-up and by applying adversarial expression consistency loss and adversarial identity consistency loss at each encoder.

3. Overview of TER-GAN

The main objective of TER-GAN is twofold: to extract descriptive, discriminative and disentangled expression and identity representations from input images, and secondly to generate synthesized expression images using the extracted expression and identity information in such away that the expression and identity information of input images are preserved. To do this, as opposed to previous facial expression

synthesis architectures, where expression or identity information is explicitly provided in the form of expression or identity codes, our TER-GAN uses two encoders to automatically extract expression and identity information. It has been reported in the literature that face images lie on a manifold [35], therefore, we argue that representing a face image with an identity code in the form of a one-hot vector is not enough to capture fine details of a target face. Similarly, in order to represent various intensities of a facial expression, using just an expression code is not sufficient to generate wide range of expression intensities. In order to learn efficient expression and identity representation, we, in addition to real images, use the synthetic expression images generated by the decoder of our generator. Thus, in this manner joint expression-invariant identity embedding and identity-invariant expression embedding are learned using two additional adversarial losses on top of representation layers at each encoder, apart from the adversarial loss imposed on the overall generator.

3.1. Network Architecture

The input to TER-GAN is a source image x_s and a target image x_t. These two images are fed to two different encoders G_{es} and G_{et}, where G_{es} aims to map source image x_s to an expression representation $f(e)$, while G_{et} is used to project the target image x_t to an identity embedding $f(i)$. The concatenation of the two embeddings: $f(x) = f(e) + f(i)$, bridges the two encoders with a decoder G_{de}. The objective of decoder G_{de} is to synthesize an expression image $\bar{x}$ having the expression e of the source image and the identity i of the target image: $\bar{x} = G_{de}(f(x))$. To further improve the quality of synthesized images, an adversarial loss is imposed on generator G by using a multi-class CNN as our discriminator D. In order to generate synthetic images with desired expressions and identities, our discriminator D performs identity and expression classification, apart from classifying between real and fake images. The overall architecture of the proposed TER-GAN is shown in Figure 2. Two additional discriminators D_{es} and D_{et} are used with G_{es} and G_{et}, respectively, to learn an identity-invariant expression embedding $f(e)$ at the FC layer of encoder G_{es} and an expression-invariant identity embedding $f(i)$ at the FC layer of encoder G_{et}. The adversarial learning scheme of identity-invariant expression embedding $f(e)$ and expression-invariant identity embedding $f(i)$ is shown with a brown-colored dashed line in in Figure 2.

3.1.1 Discriminator

The discriminator D of TER-GAN is a multi-task CNN that aims for three objectives: 1. to classify between real and fake images, 2. to classify facial expressions, and 3. to classify the identities of expression images. To achieve

these objectives, discriminator D is divided into two parts: $D = [D^e, D^i]$, where $D^e \in R^{N^e+1}$ corresponds to the part of D that is used for the classification of expressions i.e N^e denotes the number of expressions, in our case it represents six basic expressions, and an additional dimension is used to differentiate between real and fake images. Similarly, $D^i \in R^{N^i}$ is the part of D that is used to classify the identities of expression images, where N^i denotes the number of identities. The overall objective function of our discriminator D is given by the following equation:

$$\max_D \mathcal{L}_\mathcal{D}(D, G) = E_{\substack{x_s,y_s \sim p_s(x_s,y_s), \\ x_t,y_t \sim p_t(x_t,y_t)}}[\log(D^e_{y_s^e}(x_s)+$$
$$\log(D^i_{y_t^i}(x_t)]+$$
$$E_{\substack{x_s,y_s \sim p_s(x_s,y_s), \\ x_t,y_t \sim p_t(x_t,y_t)}}[\log(D^e_{N^e+1}(G(x_s,x_t))]$$
(1)

The first part of equation 1 represents the objective of D to maximize the probability of source image x_s and target image x_t to be classified to its true expression label y_s and true identity label y_t, respectively. While the second part of the function corresponds to the objective of D to maximize the probability of classifying $\bar{x}$ as a fake image.

3.1.2 Generator

In previous facial expression synthesis and manipulation methods [6], [34], [2] one encoder is used to extract feature information from an input image, while a conditional code is explicitly fed to the network to guide the facial expression synthesis process. However, in TER-GAN, the main objective of G is to efficiently extract expression and identity representation from source image x_s and target image x_t, respectively, and to generate an image $\bar{x}$ to fool D to classify it to the expression of x_s and the identity of x_t. Therefore, the generator G in TER-GAN consists of two encoders and a decoder: $G = (G_{es}, G_{et}, G_{de})$. The objective function of G is given by the following equation:

$$\max_G \mathcal{L}_\mathcal{G}(D, G) = E_{\substack{x_s,y_s \sim p_s(x_s,y_s), \\ x_t,y_t \sim p_t(x_t,y_t)}}[\log(D^e_{y_s^e}(G(x_s,x_t))+$$
$$\log(D^i_{y_t^i}(G(x_s,x_t))]$$
(2)

The adversarial loss is given as below:

$$\max_{G,D} \mathcal{L}_{adv}(D, G) = \mathcal{L}_G + \mathcal{L}_D$$
(3)

In order to improve the capability of both of our encoders to extract identity-invariant expression features and expression-invariant identity features, we introduce two additional adversarial losses on top of the representation layer

of each encoder: adversarial expression consistency loss at encoder G_{es} and adversarial identity consistency loss at G_{et}.

Encoder G_{es}: The main objective of encoder G_{es} is to extract expression representation $f(e)$ from input source image x_s. To achieve this goal, apart from employing learning by reconstruction phenomena, we propose another adversarial expression consistency loss at encoder G_{es}, which does not require any paired data and helps in learning an identity-invariant expression representation in a self-supervised manner. Specifically, since the input source image x_s and the generated image $\bar{x}$ share the same expression information but have different identities, we leverage these two images to learn an identity-invariant expression embedding. To do this, a discriminator D_{es} is trained on top of expression embedding $f(e)$, to classify the encoded features to be extracted from x_s or $\bar{x}$. To learn an identity-invariant expression embedding $f(e)$, discriminator D_{es} strives to maximize its classification accuracy, while encoder G_{es} is trained to confuse discriminator D_{es} by minimizing its accuracy. The optimization function is given by the equation below:

$$\min_{G_{es}} \max_{D_{es}} \mathcal{L}_{D_{es}} = E_{x_s \sim p_s(x_s)}, \mathcal{L}(1, D_{es}(G_{es}(x_s))) +$$
$$E_{\bar{x} \sim p_{\bar{x}}(\bar{x})} \mathcal{L}(2, D_{es}(G_{es}(\bar{x}))) \quad (4)$$

Where $\mathcal{L}$ denotes a cross-entropy loss.

Encoder G_{et}: The target image x_t is fed to encoder G_{et}, which extracts identity representation $f(i)$ for image synthesis. The target image x_t can have any expression or it can be a neutral image, since it is only used for getting identity information. Therefore, to extract an expression-invariant identity representation, we employ an adversarial identity consistency loss on top of identity representation layer $f(i)$ at encoder G_{et}. The synthesized image $\bar{x}$, which has the same identity as x_t is fed to encoder G_{et} along with the input target image x_t to learn the expression-invariant identity embedding. This goal is achieved by using a discriminator D_{et} on top of identity embedding $f(i)$, which is trained to recognize the encoded identity representation $f(i)$ as coming from x_t or $\bar{x}$. The discriminator D_{et} strives to maximize its classification accuracy while the encoder G_{et} is aimed to confuse discriminator D_{et} by minimizing its accuracy. The optimization function is given by the equation below:

$$\min_{G_{et}} \max_{D_{et}} \mathcal{L}_{D_{et}} = E_{x_t \sim p_t(x_t)}, \mathcal{L}(1, D_{et}(G_{et}(x_t))) +$$
$$E_{\bar{x} \sim p_{\bar{x}}(\bar{x})} \mathcal{L}(2, D_{et}(G_{et}(\bar{x}))) \quad (5)$$

Where $\mathcal{L}$ denotes a cross-entropy loss.

Decoder G_{de}: The input to decoder G_{de} is a concatenation of $f(e)$ and $f(i)$: $f(x) = f(e) + f(i)$, through which G_{de} will generate a synthesized image having the expression encoded in $f(e)$ and the identity information represented by $f(i)$. For this purpose, two pixel-wise reconstruction losses are used: 1. an identity reconstruction loss between input target image x_t and output image $\bar{x}$ and 2. an expression reconstruction loss between input source image x_s and output image $\bar{x}$.

$$\min_{G_{es}, G_{et}, G_{de}} \mathcal{L}_{irec} = L_1(G_{de}(G_{es}(x_s), G_{et}(x_t)), x_t) \quad (6)$$

$$\min_{G_{es}, G_{es}, G_{de}} \mathcal{L}_{erec} = L_1(G_{de}(G_{es}(x_s), G_{et}(x_t)), x_s) \quad (7)$$

However, it is known that pixel-level metrics are not very optimal for the purpose of image comparison, especially when dealing with semantics level comparison [27], [6]. Therefore, to further preserve the expression information between source image x_s and synthesized image $\bar{x}$, a pre-trained version of encoder G_{es} is used to enforce expression similarity in feature space:

$$\min_{G_{es}, G_{et}, G_{de}} \mathcal{L}_{ef} = \sum_l \omega_{1l} L_1(h_{1l}(G_{de}(G_{es}(x_s), G_{et}(x_t)), h_{1l}(x_s))$$
$$(8)$$

where h_{1l} represents the l_{th} layer feature maps extracted from the pre-trained version of G_{es} and ω_{1l} denotes the l_{th} layer's weight. The activations at all five convolutional layers of the network are used.

Similarly, to further preserve the identity information between target image x_t and synthesized image $\bar{x}$, the pre-trained version of encoder G_{et} is used to extract identity features from various inter-mediate layers. The feature-level identity preserving loss is given by the following equation:

$$\min_{G_{es}, G_{et}, G_{de}} \mathcal{L}_{if} = \sum_l \omega_{2l} L_1(h_{2l}(G_{de}(G_{es}(x_s), G_{et}(x_t)), h_{2l}(x_t))$$
$$(9)$$

3.1.3 Total Loss

The overall objective function of TER-GAN is the weighted sum of all the losses discussed in the previous sections:

$$\min_{G_{es}, G_{et}} \max_{D_{et}} \mathcal{L}_{TER-GAN} = \lambda_1 \mathcal{L}_{if} + \lambda_2 \mathcal{L}_{ef} + \lambda_3 \mathcal{L}_{irec} +$$
$$\lambda_4 \mathcal{L}_{erec} + \lambda_5 \mathcal{L}_{D_{et}} + \lambda_6 \mathcal{L}_{D_{es}} +$$
$$\lambda_7 \mathcal{L}_{adv} \quad (10)$$

3.1.4 Training the Network

Since the efficiency of learning an identity-invariant expression embedding and an expression-invariant identity embedding depends on the quality of the generated images as well, TER-GAN, having multiple loss functions, is trained
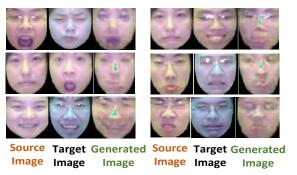
Figure 3. TER-GAN encodes expression information from source image, and identity information from target image, and generates an output expression image having the expression of source image while preserving the identity of the target image.
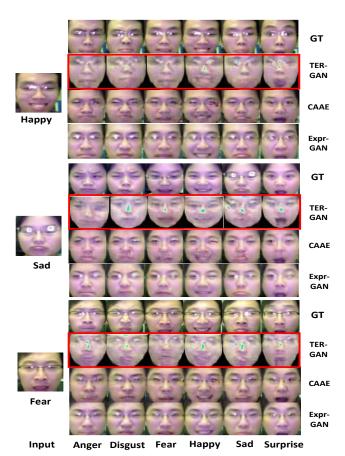
Figure 4. Comparison (visual) of facial expression editing techniques. For each input (target image), the corresponding synthetic expression images are generated. We compare our results with CAAE and Expr-GAN.

employing a curriculum strategy [6], [36] with two training stages. In the first stage of training, the encoder G_{es} is pre-trained in a supervised manner to classify facial expressions. Similarly, the encoder G_{et} is pre-trained in a fully supervised way to recognize identities. The classification layers of these two networks are then discarded and the rest of the remaining networks are attached to the over-all architecture of TER-GAN. In the second stage of training, the entire TER-GAN architecture is trained in two steps of updating the network parameters. In the first step, the expression $f(e)$ and identity $f(i)$ features are extracted from the source image x_s and the target image x_t by the two encoders G_{es} and G_{et}, and these two features are concatenated into $f(x)$ and fed to decoder G_{de} to synthesize an expression image $\bar{x}$. In the second step, the generated output image $\bar{x}$ is fed to two encoders along with their corresponding input images to further improve the quality of $f(e)$ and $f(i)$, and to learn in an adversarial manner an identity-invariant expression representation $f(e)$ and expression-invariant identity embedding $f(i)$.

The objective of training TER-GAN is to minimize equation 10. Specifically, the adversarial losses $\mathcal{L}_{D_{es}}$, $\mathcal{L}_{D_{et}}$ and $\mathcal{L}_{adv}$ lead to a min-max optimization problem, resulting in our employing a gradient reversal layer [8], [27] into TER-GAN architecture. The gradient reversal layer is implemented between D_{es} and expression embedding $f(e)$ in order to perform adversarial training scheme for $\mathcal{L}_{D_{es}}$. The gradient reversal layer does not affect the forward pass during training, but it is used to invert the gradient sign during back-propagation to practically implement the min-max training scheme. The gradient reversal layer is also used between D_{et} and identity embedding $f(i)$ in order to perform adversarial training scheme for $\mathcal{L}_{D_{et}}$.

4. Experiments

In this section we first describe the implementation details, followed by presenting the experimental results of

TER-GAN and then we demonstrate the applications of TER-GAN and its the ability to perform facial expression transfer and facial expression editing and to support facial expression recognition.

4.1. Implementation Details

The proposed technique is evaluated on the widely used Oulu-CASIA [44] dataset. Initially Convolutional Experts Constrained Local Model (CE-CLM) is used to detect facial landmarks to perform face detection and face alignment. To avoid overfitting due to using small dataset, data augmentation is performed to increase the number of images in the training dataset. From each image, five 75×75 samples are extracted from the center and four corner locations. Image rotation is then applied on each of those 75×75 cropped samples using four angles: $-6°$, $-3°$, $3°$, $6°$. Horizontal flipping is then applied on each rotated image, and thus, the size of the dataset is increased 5 times the original dataset size after data augmentation.

TER-GAN is initially pre-trained on the BU-4DFE [41]

Method	Setting	Accuracy
LBP-TOP[45]	Dynamic	68.13
HOG 3D[13]	Dynamic	70.63
STM-Explet[18]	Dynamic	74.59
Atlases[10]	Dynamic	75.52
DTAGN[11]	Dynamic	81.46
FN2EN[7]	Static	87.71
PPDN[46]	Static	84.59
DeRL[39]	Static	88.0
CNN(baseline)	Static	73.14
TER-GAN(Ours)	Static	**89.65**

Table 1. Oulu-CASIA: Accuracy for six expressions classification.

dataset, which consists of 60,600 images from 101 identities. Six image sequences are captured for each identity, and these six sequences correspond to six basic expressions. Each of these sequences are arranged in such a way that it starts from a neutral expression, reaches the peak expression in the middle, and then again ends at a neutral expression. The middle peak expression images are extracted to construct the dataset.

The pre-trained TER-GAN is then fine-tuned on the Oulu-CASIA (OC) dataset. OC dataset contains 480 video sequences captured under three different illumination conditions using two different cameras. In this experiment, only images captured under strong condition with VIS camera are used. There are 80 identities in the OC dataset, and each identity has six video sequences, corresponding to six basic expressions. Each video sequence starts with a neutral image and ends at the peak expression image. In this experiment the last three frames of each sequence are selected to construct the dataset. An identity-independent training-testing split is formed to evaluate the proposed method.

The architecture of both encoders, G_{es} and G_{et}, is designed based on five downsampling blocks consisting of a 3×3 stride 1 convolution. The number of channels are 64, 128, 256, 512, 1024 and one 30-dimensional FC layer for expression feature vector $f(e)$, and a 50-dimensional identity representation $f(i)$, constitute G_{es} and G_{et}, respectively. The decoder G_{de} is built on five upsampling blocks containing a 3×3 stride 1 convolution. The number of channels are 512, 256, 128, 64 and 3. As opposed to previous GAN architectures with a multi-task CNN based discriminator [34], [1], in TER-GAN, the discriminator D is designed in such a way that the initial downsampling convolutional layers and a FC layer are shared between D^e and D^i in order to reduce the computation cost. More specifically, four CNN blocks with 16, 32, 64, 128 channels and a 1024-dimensional FC layer are shared between the two parts. It is then divided into two branches, where, each branch has two additional FC layers with 512 and 256 channels. D^e then has an expression classification layer and D^i has an identity

classification layer. The architectures of D_{es} and D_{et} are the same, which consists of three FC layers with channels 32, 16 and 1.

TER-GAN is trained using the Adam optimizer [12], with a batch size of 64 and learning rate of 0.0002. The values of parameters are empirically set to $\omega_{11} = \omega_{21} = 0.5$, $\omega_{12} = \omega_{22} = 0.6$, $\omega_{13} = \omega_{23} = 0.7$, $\omega_{14} = \omega_{24} = 0.88$, and $\omega_{15} = \omega_{25} = 0.99$. Similarly, the weights of the total loss are set empirically as $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 1$, $\lambda_4 = 1$, $\lambda_5 = 0.3$, $\lambda_6 = 0.3$, and $\lambda_7 = 0.5$.

4.2. Facial Expression transfer

In this section, we demonstrate our model's ability to transfer facial expressions from source image x_s to target image x_t. As opposed to previous methods [6], where, a separate expression classifier is used to extract expression label in order to transfer facial expression from one image to another image, in TER-GAN, the facial expression transfer task is performed in an end-to-end manner. To do this, x_s is fed to encoder G_{es} and x_t is input to encoder G_{et} to extract expression information from x_s and identity features from x_t respectively. The expression information is then concatenated with the identity features, and the concatenated feature vector is fed to decoder G_{de} to synthesize an expression image having the expression of x_s and containing the identity of x_t. Figure 3 shows that TER-GAN transfers the facial expressions from source images quite accurately, while also preserving the target identities specified by target images.

4.3. Facial Expression Editing

In this section we demonstrate the capability of our proposed TER-GAN to edit the expression of a given image. Different from previous facial expression editing methods, like in [6], where the expression code is explicitly fed to the network, in TER-GAN, the expression information is extracted from another expression image (source image) to encode more valuable expression information than just an expression label. In this experiment, the identity feature $f(i)$ is extracted from the given image (target image, x_t) by inputting it to encoder G_{et}, while the expression information $f(e)$ is automatically extracted from another image of the same identity but with a different expression (source image x_s). $f(e)$ and $f(i)$ are then concatenated and the concatenated feature vector is then fed to the decoder G_{de} to generate a synthetic image $\bar{x}$, which has the expression information taken from x_s, while preserving the identity information. The experimental results are shown in Figure 4, where the first column corresponds to the input image (target image), while the ground truth images in top row in the right column represent the source images, x_s, which are used to extract the corresponding expression information. Although, our main objective is different than the previous
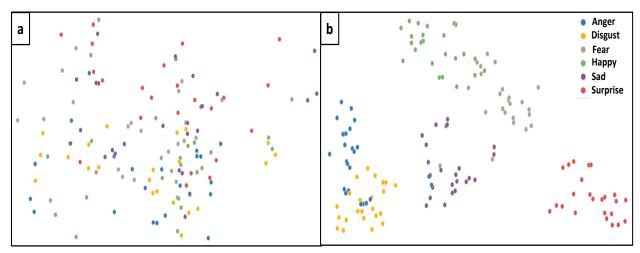
Figure 5. Expression feature space. Each color represents a different expression. Fig. 5a shows the expression distribution of features obtained from pre-trained encoder G_{es} (CNN baseline). Fig 5b. depicts the expression distribution of features obtained from encoder G_{es} after training in a TER-GAN set-up. (Best viewed in color)

methods [6], [43], and the network architecture of TER-GAN has more functionalities, and is more complex than the models proposed in [6] and [43], we, for the sake of comparison, compare our results with [6] and [43]. As it can be seen in Figure 4, our TER-GAN can not only synthesize an image of the desired expression, but the identity information is preserved more in our case than in [6] and [43]. To reduce the computational cost and model complexity, we have not used any variation regularization [20] method as used in [6] to reduce spike artifacts on the reconstructed images, which, we believe, if incorporated in TER-GAN, will enhance the visual quality of our synthesized images as well.

4.4. Facial Expression Recognition

In this section we demonstrate the ability of TER-GAN to efficiently disentangle the expression information of any expression image from its identity information. One of the major issues with conventional FER techniques is that the representation used for facial expression recognition contains identity information as well as the expression information and, as a result, the performance of FER degrades on unseen identities during real-time applications. Therefore, in order to obtain identity free expression information, the encoder G_{es} of TER-GAN is detached from the rest of the architecture after training and is used to perform facial expression recognition. Specifically, an expression image is fed to the detached encoder G_{es} and the output expression representation $f(e)$ is extracted. This feature vector is then fed to a shallow classifier for facial expression recognition.

To evaluate the performance of the proposed disentangled facial expression recognition technique, we conducted an eight fold cross validation on the Oulu-CASIA dataset. Table 1 shows the average accuracy obtained us-

ing the proposed technique. The reported results show that TER-GAN outperforms state-of-the-art techniques including GAN-based methods like DeRL [39] and deep CNN-based techniques like DTAGN-Joint [11], FN2EN [7], and PPDN [46]. Although, we are using only images to extract expression information for FER, our method out-performs techniques like DTAGN[11], Atlases[10], STM-Explet[18], HOG 3D[13] and LBP-TOP[45] that exploit temporal information of video sequences.

4.5. Expression Feature Visualization

In this part, we demonstrate that the expression representation $f(e)$ learned by our proposed TER-GAN is disentangled from identity information. To do this, we first extract the expression feature vector $f(e)$ from encoder G_{es}, and then employ t-SNE [19] to project the 30-dim feature vector $f(e)$ on a two dimensional space for visualization purpose. The 2d expression feature space is shown in Figure 5. For the sake of comparison with our CNN baseline, we conduct two experiments to show that the expression representation learned by encoder G_{es}, when trained in a TER-GAN set-up, is disentangled from identity information. In the first experiment, we extract 30-dim expression features from our CNN baseline network (pre-trained encoder G_{es}), and visualize this in a 2d feature space using t-SNE. The result of our first experiment is shown in Figure 5(a). It can be seen that the expression features are all entangled with each other in the expression feature space, which clearly indicates that the CNN baseline model fails to disentangle expression information from identity features. In the second experiment, we employ encoder G_{es} trained in an end-to-end manner in a TER-GAN set-up, to extract expression representation $f(e)$, and project it to 2d feature space using t-SNE. The result of the second experiment is shown in Fig-

ure 5(b). We can see that the expression features are organised in the form of six clusters, corresponding to six basic expressions, which indicates that the expression information is effectively disentangled from identity information.

5. Conclusion

In this paper we have proposed a unified expression transfer, editing and recognition architecture, TER-GAN, which has two objectives: 1). to extract efficient and disentangled expression and identity features from input images, and 2). to employ the extracted expression and identity representations for realistic looking expression synthesis that preserves the identity information of the target (given) image. This goal is achieved by explicitly encoding the expression information from a source image and extracting identity information from a target image by using two different dedicated encoders, and these two feature vectors are than combined to generate an expression image by employing the decoder part of TER-GAN. In order to further improve the expression and identity feature extraction process, we have introduced novel expression and identity consistency losses. Experimental results show that the proposed method can be used for efficient facial expression transfer and facial expression editing, and the disentangled feature representation can be used for facial expression recognition.

References

[1] Kamran Ali and Charles E Hughes. Facial expression recognition using disentangled adversarial learning. *arXiv preprint arXiv:1909.13135*, 2019. 1, 2, 7

[2] Kamran Ali, Ilkin Isler, and Charles Hughes. Facial expression recognition using human to animated-character expression translation. *arXiv preprint arXiv:1910.05595*, 2019. 4

[3] Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*, 2014. 1

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul 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*, pages 8789–8797, 2018. 1

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul 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*, pages 8789–8797, 2018. 3

[6] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[7] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 118–126. IEEE, 2017. 7, 8

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 6

[9] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. Automatic face reenactment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4217–4224, 2014. 3

[10] Yimo Guo, Guoying Zhao, and Matti Pietikäinen. Dynamic facial expression recognition using longitudinal facial expression atlases. In *European Conference on Computer Vision*, pages 631–644. Springer, 2012. 7, 8

[11] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991, 2015. 7, 8

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[13] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. 2008. 7, 8

[14] Ying-Hsiu Lai and Shang-Hong Lai. Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 263–270. IEEE, 2018. 3

[15] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2018. 3

[16] Mu Li, Wangmeng Zuo, and David Zhang. Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586*, 2016. 3

[17] Alexandra Lindt, Pablo Barros, Henrique Siqueira, and Stefan Wermter. Facial expression editing with continuous emotion labels. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. 1, 2

[18] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756, 2014. 7, 8

[19] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8

[20] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015. 8

[21] Umar Mohammed, Simon JD Prince, and Jan Kautz. Visiolization: generating novel facial images. *ACM Transactions on Graphics (TOG)*, 28(3):57, 2009. 1

[22] Naima Otberdout, Mohamed Daoudi, Anis Kacem, Lahoucine Ballihi, and Stefano Berretti. Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets. *arXiv preprint arXiv:1907.10087*, 2019. 1

[23] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2018. 2

[24] F Qiao, N Yao, Z Jiao, Z Li, H Chen, and H Wang. Geometry-contrastive generative adversarial network for facial expression synthesis. corr abs/1802.01822 (2018), 1802. 3

[25] Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, and Hongan Wang. Emotional facial expression transfer from a single image via generative adversarial nets. *Computer Animation and Virtual Worlds*, 29(3-4):e1819, 2018. 2

[26] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, pages 1431–1439, 2014. 1

[27] Amélie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Mosseri, Forrester Cole, and Kevin Murphy. Xgan: Unsupervised image-to-image translation for many-to-many mappings. *arXiv preprint arXiv:1711.05139*, 2017. 5, 6

[28] Zhiwen Shao, Hengliang Zhu, Junshu Tang, Xuequan Lu, and Lizhuang Ma. Explicit facial expression transfer via fine-grained semantic representations. *arXiv preprint arXiv:1909.02967*, 2019. 1, 2

[29] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 650–665, 2018. 3

[30] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 627–635. ACM, 2018. 2, 3

[31] Joshua M Susskind, Geoffrey E Hinton, Javier R Movellan, and Adam K Anderson. Generating facial expressions with deep belief nets. In *Affective computing*. IntechOpen, 2008. 1

[32] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 1

[33] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424, 2017. 3

[34] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424, 2017. 4, 7

[35] Xueping Wang, Weixin Li, Guodong Mu, Di Huang, and Yunhong Wang. Facial expression synthesis by u-net conditional generative adversarial networks. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 283–290. ACM, 2018. 2, 4

[36] Olivia Wiles, A Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–686, 2018. 2, 6

[37] Fei Yang, Lubomir Bourdev, Eli Shechtman, Jue Wang, and Dimitris Metaxas. Facial expression editing in video using a temporally-smooth factorization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 861–868. IEEE, 2012. 3

[38] Fei Yang, Jue Wang, Eli Shechtman, Lubomir Bourdev, and Dimitri Metaxas. Expression flow for 3d-aware face component transfer. *ACM transactions on graphics (TOG)*, 30(4):60, 2011. 1, 3

[39] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2168–2177, 2018. 7, 8

[40] Raymond Yeh, Ziwei Liu, Dan B Goldman, and Aseem Agarwala. Semantic facial expression editing using autoencoded flow. *arXiv preprint arXiv:1611.09961*, 2016. 1

[41] L Yin, X Chenand Y Sun, T Worm, and M Reale. A high-resolution 3d dynamic facial expression database, 2008. In *IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam, The Netherlands*, volume 126. 6

[42] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3359–3368, 2018. 3

[43] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5810–5818, 2017. 8

[44] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. 6

[45] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):915–928, 2007. 7, 8

[46] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *European conference on computer vision*, pages 425–442. Springer, 2016. 7, 8