2018 13th IEEE International Conference on Automatic Face & Gesture Recognition

Emotion-Preserving Representation Learning via Generative Adversarial Network for Multi-view Facial Expression Recognition

Ying-Hsiu Lai and Shang-Hong Lai
Dept. of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
Email: lai@cs.nthu.edu.tw

***Abstract*—Face frontalization is one way to overcome the pose variation problem, which simplifies multi-view recognition into one canonical-view recognition. This paper presents a multi-task learning approach based on the generative adversarial network (GAN) that learns the emotion-preserving representations in the face frontalization framework. Taking advantage of adversarial relationship between the generator and the discriminator in GAN, the generator can frontalize input non-frontal face images into frontal face images while preserving the identity and expression characteristics; in the meantime, it can employ the learnt emotion-preserving representations to predict the expression class label from the input face. The proposed network is optimized by combining both synthesis and classification objective functions to make the learnt representations generative and discriminative simultaneously. Experimental results demonstrate that the proposed face frontalization system is very effective for expression recognition with large head pose variations.**

***Keywords*-face frontalization; facial expression recognition; pose variation;**

I. Introduction

Recently, due to the emergence of deep learning, significant progress has been made in face-related tasks. Mollahossein *et al.* [12] proposed a deep neural network architecture inspired by GoogLeNet [19] and AlexNet [8] for developing a facial expression recognition system, which outperforms traditional methods based on handcrafted features. Jung *et al.* [7] joint fine-tuned two small deep network models, temporal appearance network and temporal geometry network, to obtain more discriminative features and achieved higher performance on two public facial expression recognition databases, i.e., CK+ database [10], Oulu-CASIA database [25]. However, most of the facial expression recognition methods focus on analysis of expressions from frontal faces. Even applying deep learning technologies has made significant improvements, pose variation is still a challenging problem for many realistic face-related applications.

To deal with the problem of head pose variations, increasing massive face images with arbitrary views for training is a common and easy way to learn pose-robust representations. However, collecting and labeling a large number of face images is quite a huge work, and the improvement of recognition is limited. Another intuitive approach is to simplify the problem of face-related recognition under large pose variations by reducing it to the canonical view recognition, i.e., automatic synthesis of the corresponding frontal face image from a non-frontal face image.

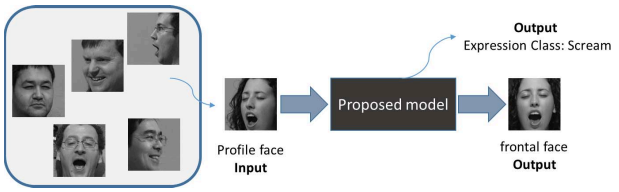


Figure 1. The flow chart of the proposed method for multi-task learning. Given a non-frontal face image, the proposed model would predict its expression and synthesize its frontal view at the same time.

Frontalization is to synthesize a frontal face image from a non-frontal face image. By using deep learning algorithms, Zhu *et al.* [28] first developed a simple neural network to learn identity-preserving pose-invariant features from a frontalization process, and they obtained great improvement for non-frontal face recognition. After that, some other face recognition researches [21], [20], [5] also presented more complicated deep models, e.g., convolutional neural network (CNN), generative adversarial network [2], and obtained better synthesis results and better handled large pose variations. However, the existing frontalization approaches can only preserve face identity with neural expression, but they cannot preserve facial expressions after the frontalization. Therefore, we aim to develop a face frontalization system that can preserve the facial expression in this paper.

This paper presents a multi-task learning via generative adversarial networks for multi-view facial expression recognition. As the example shown in Fig.1, given a profile face image at an arbitrary head pose and with an arbitrary expression, the proposed model would generate two kinds of outputs: the expression class label and synthesized frontal face image. We design different kinds of objective functions for learning the emotion-preserving representations during the frontalization process, which can not only facilitate the synthesized frontal face image maintaining more expression characteristics, but also obtain more discriminative pose-invariant features for the expression recognition under large poses.

978-1-5386-2335-0/18/$31.00 ©2018 IEEE
DOI 10.1109/FG.2018.00046

II. Related Works

A. Facial Expression Recognition

For deep learning based approaches, the facial expression recognition systems have been developed very well for frontal face expressions. Mollahossein *et al.* [12] exploited deep neural network architecture to simplify traditional hand-crafted feature extraction and feature selection methods, but these methods still outperformed the traditional handcraft-based methods. In addition, Jung *et al.* [7] developed jointly fine-tuned small deep network models, temporal appearance model and temporal geometric model, which provide better recognition results than the single deeper network model. However, the existing methods still suffer from the accuracy degradation under large head pose variations.

To overcome the head pose variation problem, facial expression recognition methods can be grouped into two categories: 3D-based and 2D-based approaches. 3D-based methods typically exploit 3D features or map the 3D data onto a representation [18]. Since 3D face data is information-rich by nature, 3D-based methods make them more robust to pose variation. However, 2D-based studies are used more often in practice, because 2D data can be easily obtained and processed with different ways. For 2D-based methods, the researchers usually focused on developing discriminative pose-invariant features or handling facial expression recognition separately on different face views. Zhang *et al.* [24] proposed a deep neural network (DNN) model to learn the relationship between extracted low-level SIFT features and high-level information. Jampour *et al.* [6] introduced a mapping algorithm that maps the features extracted from non-frontal view to an approximately frontal view feature space according to the head pose estimation.

B. Face Frontalization

Face frontalization means automatically synthesizing a frontal face image from a face image at an arbitrary head pose. There are two ways to accomplish face frontalization: 3D transformation solutions and 2D deep learning-based solutions.

In 3D transformation solutions, Hassner *et al.* [4] aligned 2D non-frontal face to 3D reference model points by utilizing facial landmarks, and then computed projection matrix to transform the non-frontal face into frontal view. Zhu *et al.* [27] not only used facial landmark to align with 3D points, but also meshed 2D face into 3D object and normalized facial expression during the frontalization process.

In contrast, 2D deep learning-based solutions do not need to design algorithms in each step manually; they just designed a deep network architecture to learn the whole process of frontalization directly. The first approach [28] presented the possibility of utilizing deep models to learn identity-preserving representation during the frontalization. Yim *et al.* [21] went on to develop a novel network that can not only generate frontal face views, but also generate arbitrary face views of the desired poses. At the same time, they designed a new multi-task learning strategy that can recover the generated face back to original face views, in order to improve the identity preserving ability.

Furthermore, [20], [5] showed that the frontalization results can be greatly improved when generative adversarial networks (GAN) [2] replaced the simple deep neural networks, and the adversarial loss replaced the simple L_2 loss. Especially the TP-GAN network proposed by Huang *et al.* [5] can even deal with 90-degree large pose frontalization. However, no deep learning-based approaches have put their interests on preserving facial expression characteristics during the frontalization, since they only focused on improving the performance in face recognition.

III. Proposed Method

This paper proposes a novel facial expression representation learning method based on GAN network, which can not only recognize expressions but also synthesize the corresponding frontal face images at the same time. We aim to synthesize the frontal face images from profile face images and learn the emotion-preserving representation under the face frontalization process. Our goal is to train a multi-task learning network that can generate the emotion-preserving frontal views to achieve canonical-view facial expression recognition, and simultaneously recognize expressions in our network model. Therefore, the proposed GANmodel as depicted in Figure 2 is not only for improving the face synthesis quality, but also for better representation learning on both the generator and the discriminator due to the adversarial loss in GAN.

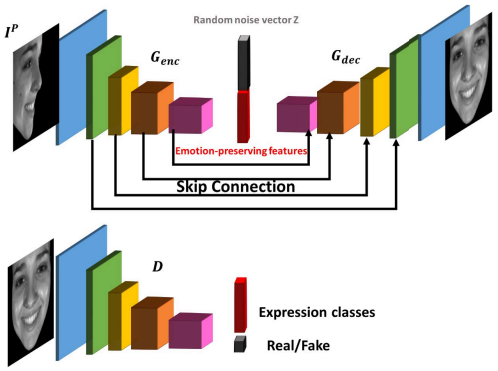


Figure 2. The proposed GAN model

A. Network Architecture

Given a pair of corresponding face images $\{I^P, I^F\}$, with expression class label y, where I^P is a profile face with arbitrary head poses, and I^F is the corresponding frontal face image with its identity and expression the same as those of the profile image I^P. The proposed GAN model contains an

encoder-decoder structure generator G and a discriminator D.

Inspired by the existing GAN-based face frontalization approaches [20], [5] for face recognition fields, our generator G is designed with an encoder-decoder structure model, consisting of encoder G_{enc} and decoder G_{dec}, where G_{enc} was taken to learn the emotion-preserving representation from I^P, and G_{dec} was taken to recover the frontal view of I^P as similar as I^F from the extracted features. Moreover, the bottleneck of G_{enc}, the extracted emotion-preserving features, can be used to recognize expressions directly. Therefore, we apply an additional fully-connection layer at the end of the bottleneck layer in G_{enc} to enforce G be trained for multi-task learning.

For the discriminator D, its main goal is to distinguish the real images I^F from the fake generated images $G(I^P)$. The minimax two-player game lets the generated frontal faces $G(I^P)$ move towards the same distribution as real images I^F, and makes it really difficult to separate the generated images from the real images. Furthermore, D as a discriminator, of course it also can be trained to recognize expression at the same time. Not only the generative model benefits from the adversarial relationship between G and D, but also the representation learning ability, so the design of multi-task learning for D can make G learn more discriminative emotion-preserving representations and improve the performance for facial expression recognition.

The detailed structures of D and G_{enc} are provided in Table I. Take Conv1 layer for example, Conv1 is a convolutional layer with filter size 33, stride-1, and its outputs are 32 128128 feature maps, and so on. In addition, FC means fully connected layer. In D and G_{enc}, we replace common pooling layers with 2-strided convolutions. Particularly, there are two branch layers on top of the Conv5 layer. Branch Conv6 is for facial expression recognition, and branch Conv9 is for distinguishing fake generated images from real images. Therefore, both D and G_{enc} consist of branch Conv6 for expression recognition, but only D includes branch Conv9 to judge real/fake. Specifically, Conv8 layer represents the to-be-learned emotion-preserving features from G_{enc}.

In the proposed GAN model, we use a random vector to represent some other face variations, except expressions, for the face synthesis. However, in our experiments without adding the noise vector, the expression recognition result is very similar to that with adding the noise vector.

Table I
NETWORK ARCHITECTURE FOR D AND G_{enc}.

Layer	Filter Size	Output Size
Conv1	$3\times3/1$	$128\times128\times32$
Conv21	$3\times3/2$	$64\times64\times64$
Conv22	$3\times3/1$	$64\times64\times64$
Conv23	$3\times3/1$	$64\times64\times64$
Conv31	$3\times3/2$	$32\times32\times128$
Conv32	$3\times3/1$	$32\times32\times128$
Conv33	$3\times3/1$	$32\times32\times128$
Conv41	$3\times3/2$	$16\times16\times256$
Conv42	$3\times3/1$	$16\times16\times256$
Conv43	$3\times3/1$	$16\times16\times256$
Conv5	$3\times3/2$	$8\times8\times512$
Conv6	$3\times3/2$	$4\times4\times512$
Conv7	$3\times3/2$	$2\times2\times512$
Conv8	$3\times3/2$	$1\times1\times256$
FC1	–	class number for D and G_{enc}
Conv9	$3\times3/4$	$2\times2\times256$ for D only
FC2	–	1(real/fake) for D only

B. Objective Function

This paper adopts different loss functions to optimize the proposed model, the optimization losses can be grouped into two categories: synthesis loss and classification loss. The following sections will describe each individual loss function included in the total loss function in detail.

1) Adversarial Loss: The basic network structure of GAN [2] contains a generator G and a discriminator D that compete with two-player minimax game. D tries to distinguish the real frontal face images I^F from fake generated frontal faces $G(I^P)$, and G tries to generate realistic-like frontal faces to fool D. The corresponding adversarial losses are listed below, D is trained to maximize L_{ad_D}, and G is trained to minimize L_{ad_G}:

$$\begin{aligned} L_{ad\_D} &= E[\log(D(I^F))] + E[1-\log(D(G(I^P)))] \\ L_{ad\_G} &= -E[\log(D(G(I^P)))] \end{aligned} \tag{1}$$

As the minimax game formulation introduced in [2], originally G was optimized by minimizing $(1-\log?D(G(I^P)))$. However, since D converges early in learning that G cannot obtain sufficient gradients to learn well. So it is better for G to alternatively maximize $\log?D(G(I^P))$ (same as minimizing $-\log D(G(I^P))$) in practice. Adversarial loss makes the synthesis images look like the real frontal face images. It can prevent blurred effects and synthesize high-fidelity images.

2) Pixel-wise Loss: To speed up the convergence of G and facilitate the image content consistency, we adopt pixel-wise L_1 loss between the synthesized frontal faces $G(I^P)$ and ground truth frontal faces I^F, given by:

$$L_{pixel} = \frac{1}{W\times H}\sum_{x=1}^{W}\sum_{y=1}^{H}|G(I^P_{x,y}) - I^F_{x,y}| \tag{2}$$

where W,H represent the width and height of the image, and x,y means the position in the image space.

Instead of using L_2 loss but L_1 loss is because L_1 loss is more robust, and L_2 loss is too sensitive to the training samples and easily influenced by "outlier". Although pixel-wise loss would cause blurred effects, it is still an important

part for accelerating the optimization speeds and improving synthesis performance.

3) Symmetry Loss: Due to the self-occlusion on profile faces, it is quite hard to recover the occluded facial parts back to frontal face views. Generally, human faces have symmetrical characteristics that the left and right sides of face are bilateral symmetry. Therefore, we exploit the symmetry traits of human face as a prior to solve the self-occlusion problem slightly on large pose cases and thus may improve the frontal face synthesis results. The equation of symmetry loss is given as follows:

$$L_{sym} = \frac{1}{(W/2) \times H} \sum_{x=1}^{W/2} \sum_{y=1}^{H} |G(I^P_{x,y}) - G(I^P_{W-(x-1),y})| \quad (3)$$

However, human faces, especially with expressions, are not symmetric all the time, so we adjust the weighting for symmetry loss to reduce the symmetry constraint for face image synthesis.

4) Feature Loss: In the Improved-GAN [17], they introduced many kinds of improved techniques for training the GAN model; the feature loss L_{feat} actually is one of the improved techniques called feature matching. Feature matching facilitates the generator to generate the images that match the probability distribution of real frontal face images, which is a way to prevent the generator G from overtraining on the current discriminator. The feature loss function is given by

$$L_{feat} = \frac{1}{N} \sum_{i=1}^{N} |F(I^P_i) - F(I^F_i)| \quad (4)$$

where F represents the features for matching, and N is the total number of the features.

Originally, improved-GAN trained the generator by L_{feat} on an intermediate layer of the discriminator, whereas we optimize L_{feat} on emotion-preserving feature vector in G_{enc}. We compute the feature loss by L_1 loss between the features of profile images I^P and the features of truth frontal images I^F, in order to obtain more discriminative features, in addition to match the probability distribution of real frontal face images.

5) Classification Loss: Besides the synthesis optimization functions, we adopt the classification loss to optimize the performance of facial expression recognition on both G and D. According to auxiliary classifier GAN (AC-GAN) [14], every training sample has a class label y, and the discriminator D estimates both the real frontal face probability distribution and the class label probability distribution. In other words, D is optimized by the log-likelihood of the real images L_{ad_D} (Section III-B1) and the log-likelihood of the correct class L_{class} (Equation 5). L_{class} includes the log-likelihood of correct class on both real frontal faces $D(I^F)$ and synthesis frontal faces $D(G(I^P))$.

$$L_{class} = E[\log(D(I^F) = y] + E[\log(D(G(I^P)) = y)] \quad (5)$$

$$L_{class\_enc} = E[\log(G_{enc}(I^P) = y)] \quad (6)$$

For the same situation, G is also trained to optimize L_{class} to match real frontal faces probability distribution and at once learn more discriminative emotion-preserving representation. Moreover, the proposed G is trained with an additional classification loss L_{class_enc} (Equation 6) to minimize the log-likelihood of the correct class from the emotion-preserving features in G_{enc} directly, which can make G be able to deal with multi-task learning.

6) Overall Objection Function: To sum up, the overall objective function for D is denoted by L_{θ_D}, and the overall objective function for G is denoted by L_{θ_G}, and they are given as follows:

$$L_{\theta_D} = \mu_1 L_{ad\_D} + \mu_2 L_{class}, \quad (7)$$

$$L_{\theta_G} = L_{synthesis} + L_{classification}, \quad (8)$$

$$L_{synthesis} = \lambda_1 L_{ad\_G} + \lambda_2 L_{pixel} + \lambda_3 L_{sym} + \lambda_4 L_{feat}, \quad (9)$$

$$L_{classification} = \lambda_5 L_{class\_enc} + \lambda_6 L_{class}, \quad (10)$$

where μ's and λ's are parameters for adjusting the weights for individual loss functions.

The synthesis loss in L_{θ_G} includes adversarial loss, pixel-wise loss, symmetry loss, and feature loss. The classification loss in L_{θ_G} consists of L_{class_enc} and L_{class}.

IV. Experimental Evaluation

The proposed method aims at both representation learning and frontal face synthesis. We quantitatively demonstrate the representation learning capability of our method for multi-view facial expression recognition in Sec IV-B, and illustrate the qualitative frontal face synthesis results in Sec IV-C.

A. Experimental Setting

1) Implementation Details: In pre-processing, we apply MTCNN [23] to detect human face. According to the predicted bounding box and five facial landmark points, we crop the detected face and resize it into a 128×128 grayscale image, with a setting that the nose would be the center in x-axis coordinates (left-right). Our network is implemented on Tensorflow [1]. We use Adam optimizer with learning rate of 10^{-4} and momentum of 0.5. We empirically set the weighting parameters $\mu_1 = 0.5$, $\mu_2 = 0.5$, $\lambda_1 = 10^{-3}$, $\lambda_2 = 1$, $\lambda_3 = 0.3$, $\lambda_4 = 0.03$, $\lambda_5 = 0.1$, $\lambda_6 = 0.05$ for all experiments. Batch size is set to 78 in Multi-PIE database, 60 in BU-3DFE database.

Table II
COMPARISON WITH EXISTING METHODS IN MULTI-PIE DATABASE.

Methods	Feature	Subjects	Pose	Expression number	Accuracy
Moore [13]	LBP^{ms}	100, 10-fold	7	6	73.98%
Moore [13]	*LGBP*	100, 10-fold	7	6	80.17%
GSRRR [26]	LBP^{u2}	100, 5-fold	7	6	81.7%
2D JFDNN [7]	Image+landmarks(DNN)	100, 5-fold	7	6	82.9%
Zhang [24]	SIFT(DNN)	100, 5-fold	7	6	82.0%
KPSNM [6]	HoG+LBP	145, X	13	6	82.55%
KPSNM [6]	HoG+LBP	145, X	7	6	83.09%
Ours $G_{enc}(I^P)$	Image(GAN)	100, 5-fold	13	6	**86.74%**
Ours $D(G_{enc}(I^P))$	Image(GAN)	100, 5-fold	13	6	**87.08%**

Table III
OVERALL ACCURACIES WITH RESPECT TO DIFFERENT VIEWPOINTS IN MULTI-PIE DATABASE.

Acc. (%)	0°	15°	30°	45°	60°	75°	90°	Avg.
$D(G(I^P))$	90	91.17	89.42	89.92	89	85.83	75.75	**87.09**
$G_{enc}(I^P)$	90	89.75	90.08	88.67	87.5	85.33	77.58	**86.76**

Table IV
CORRESPONDING ACCURACIES ON DIFFERENT EXPRESSIONS AT DIFFERENT HEAD POSE ANGLES IN MULTI-PIE DATABASE.

Acc. (%)	0°	15°	30°	45°	60°	75°	90°	Avg.
Neutral	92	**93**	92.5	89.5	90.5	88	70.5	87.69
Smile	95	98.5	**99**	98	95.5	94.5	83.5	94.85
Squint	**80**	78.5	78	77.5	74.5	80	74.5	77.38
Surprise	**98**	96	96.5	97	96.5	90	81.5	93.31
Disgust	**78**	76	78	74.5	73	67.5	69.5	73.46
Scream	**97**	96.5	96.5	95.5	95	92	86	93.85
Overall	90	89.75	**90.08**	88.67	87.5	85.33	77.58	86.76

2) Databases: The Multi-PIE database [3] contains more than 750000 images of 337 subjects with four recording sessions, in which 235 are male, 107 are female. Subjects were recorded with 15 cameras at different viewpoints and 19 different illumination conditions. In addition, subjects were asked to perform different expressions in each session. For each recording session, the participants and recorded expressions are a bit different. In total, there are six kinds of expressions recorded, which consist of neutral, smile, squint, surprise, disgust, and scream. Although there are 337 subjects in the Multi-PIE database, only 100 subjects presented in all four recording sessions are selected, in order to balance the training data class labels. 13 face views are chosen in the experiments, i.e., 0°, ±15°, ±30°, ±45°, ±60°, ±75°, ±90° face views are considered.

The BU-3DFE database [22] contains 100 subjects (56% female, 44% male) with a variety of ethnics, ages. Every subject was recorded with six standard facial expressions, i.e., anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), and surprise (SU), of four levels of intensities. Therefore, there are 24 instant 3D expression models for each subject. To study on the multi-view facial expression recognition fields, we render these 3D expression models and adjust viewpoints to project them into 2D face images with specified head poses, i.e., with 0°, ±30°, ±45°, ±60°, ±90° yaw angles.

B. Representation Learning

1) Results on Multi-PIE Database: We compare our method with the state-of-the-art methods for multi-view facial expression recognition in Multi-PIE database. We use 5-fold cross validation for the experiments, *i.e.*, we randomly divide 100 subjects into 80 subjects for training and 20 subjects for testing, and there are no overlap between the training subjects and the testing subjects.

Table II shows the comparison between our work and some previous methods. Among these methods, [7], [24] are deep learning based approaches and the others are traditional methods using hand-crafted features. Instead of taking 7 poses (0°, 15°, 30°, 45°, 60°, 75°, 90° yaw angles) which were used in most of the previous works, we use 13 poses (0°, ±15°, ±30°, ±45°, ±60°, ±75°, ±90° yaw angles) to conduct the experiment, which lead to worse performance according to the results of [6]. However, our method outperforms the state-of-the-art methods in this experiment. $G_{enc}(I^P)$ represents the performance of expression recognition based on the encoder G_{enc}, and $D(G(I^P))$ represents the result that D uses the synthesis image $G(I^P)$ to recognize expressions. As the result, our generated frontal face images have preserved the expression characteristics that are effective for the recognition task.

Table III shows the overall accuracies of both $G_{enc}(I^P)$ and $D(G(I^P))$ under different head poses, and Table IV presents the corresponding accuracies on every expression under different head poses predicted by G_{enc}. Consequently, the performance of $D(G(I^P))$ is very similar to $G_{enc}(I^P)$. From Table IV, most of the expressions are easier to recognize under smaller pose angles.

Table V
COMPARISON OF DIFFERENT EXPRESSION RECOGNITION METHODS ON BU-3DFE DATABASE.

Methods	Feature	Subjects	Pose	Expression number	Accuracy
Moore [13]	LBP^{ms}	100, 10-fold	5	6(4 levels)	65.0%
Moore [13]	$LGBP$	100, 10-fold	5	6(4 levels)	68.0%
GSRRR [26]	LBP^{u2}	100, 5-fold	5	6(4 levels)	66.0%
2D JFDNN [7]	Image+landmarks(DNN)	100, 5-fold	5	6(4 levels)	72.5%
Zhang [24]	SIFT(DNN)	100, 5-fold	5	6(4 levels)	80.1%
Ours $G_{enc}(I^P)$	Image(GAN)	100, 5-fold	5	6(4 levels)	**73.13%**

2) Ablation Study: To demonstrate the contribution of the loss function proposed in this paper to the final expression recognition accuracy, we perform an ablation study to evaluate the model accuracies by incrementally adding the loss terms. Here, we denote the adversarial loss, pixel-wise loss, symmetry loss, feature loss, and classification loss by L_{ad}, L_{pixel}, L_{sym}, L_{feat}, and $L_{classification}$, respectively. The detailed definitions of these loss terms are defined in the previous section. The expression classification accuracies by using the proposed deep network model trained with different combinations of the loss terms under the same experimental setting for MULTI-PIE dataset are listed in Table VI. It is clear to see the incremental improvement of recognition accuracies by adding the loss terms.

Table VI
RECOGNITION ACCURACIES FOR THE PROPOSED MODEL TRAINED WITH DIFFERENT COMBINATIONS OF LOSS TERMS FOR THE EXPERIMENTS ON MULTI-PIE DATABASE.

Loss terms	$D(G(I^P))$	$G_{enc}(I^P)$
L_{pixel}	81.71	N/A
$L_{pixel}+L_{ad}$	81.18	N/A
$L_{pixel}+L_{ad}+L_{sym}$	82.21	N/A
$L_{pixel}+L_{ad}+L_{sym}+L_{feat}$	82.66	N/A
$L_{pixel}+L_{ad}+L_{sym}+L_{feat}+L_{classification}$	87.09	86.76

3) Results on BU-3DFE Database: Similar to the setting in Multi-PIE database, we apply 5-fold cross validation, and take the 5 head poses (0°, 30°, 45°, 60°, 90°) for the experiment in BU-3DFE database. We also use the images with four levels of expression intensities to conduct the experiment. It is challenging for the generator to preserve much detail of face characteristics (e.g. identity, expression) and learn discriminative features for expression recognition.

Table V shows the comparison with other previous methods on BU-3DFE database. Our method outperforms most of the previous works except [24], which is one of the deep learning based approach that utilized SIFT features as their designed deep neural network (DNN) inputs to learn higher-level representations. This is probably because the proposed method did not explicitly use detailed local expression representation, thus making the proposed model unable to achieve the best expression recognition accuracy for the BU-3DFE database. However, the proposed model still can compete with another deep learning based approach [7] that utilized image feature and landmarks to jointly train a DNN discriminator.

Table VII
OVERALL ACCURACIES WITH RESPECT TO DIFFERENT VIEWPOINTS IN BU-3DFE DATABASE.

Acc. (%)	0°	30°	45°	60°	90°	Avg.
$G_{enc}(I^P)$	74.25	74.00	73.38	73.29	70.75	**73.13**

In Comparison with the results on the Multi-PIE database, the expressions in BU-3DFE are more difficult to recognize, even by humans. For example, the difference between Fear, Sadness, and Anger are very subtle and these expressions are easily confused with each other. For the experiments on both BU-3DFE database and Multi-PIE database, the competitive recognition performance by using $D(G(I^P))$ indicates that the synthesized frontal face images by using our proposed network are effective for the expression recognition task.

C. Face Synthesis

1) Synthesis Results on Databases: Table VIII and Table IX show the frontal face synthesis results on Multi-PIE database and BU-3DFE database (more synthesis results are shown in Supplement). Consequently, the learnt representations facilitate the synthesized frontal face images to preserve a certain expression, identity characteristics. In addition, the advantage of GAN makes the synthesized faces similar to the corresponding real frontal face I^F and achieves high-quality image synthesis results, even in the cases with large head poses. Specifically, the details of facial characteristics like wrinkles and beards are difficult for our model to reconstruct perfectly, but the learnt emotion-preserving representations make the generator reconstruct the corresponding frontal faces with expressions.

2) Comparison of Frontalization Results: To further demonstrate the image synthesis ability of our work, we conduct an experiment to compare the synthesis results with other existing face frontalization methods. For the consideration that some methods can only handle the pose smaller than 45 angle, we present the frontalization results that are under small head poses (Table X). In fact, our work can synthesize realistic-looking frontal faces under very large poses.

Table VIII
SYNTHESIS RESULTS ON MULTI-PIE DATABASE: ROW (A) SHOWS THE INPUT PROFILE IMAGES I^P, AND ROW (B) SHOWS THE SYNTHESIS RESULTS $G(I^P)$ FOR DIFFERENT EXPRESSIONS. THE INPUT IMAGES UNDER 0° ANGLE CAN BE CONSIDERED AS THE REAL FRONTAL FACE IMAGES I^F. EACH COLUMN REPRESENTS THE CORRESPONDING VIEWPOINTS: $0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 90^\circ$ YAW ANGLES.

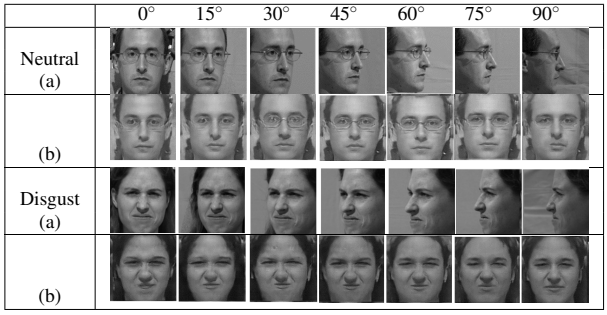
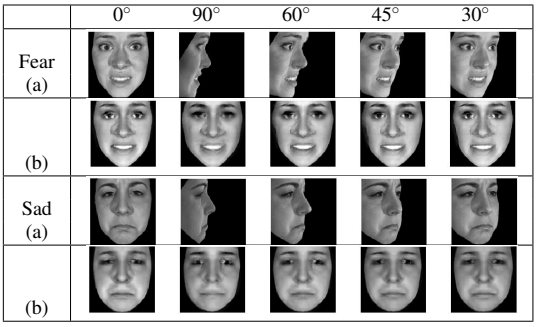


Table IX
SYNTHESIS RESULTS ON 3D-BUFE DATABASE: ROW (A) SHOWS THE INPUT PROFILE IMAGES I^P, AND ROW (B) SHOWS THE SYNTHESIS RESULTS $G(I^P)$ FOR DIFFERENT EXPRESSIONS. THE INPUT IMAGES UNDER 0° ANGLE CAN BE CONSIDERED AS THE REAL FRONTAL FACE IMAGES I^F. EACH COLUMN REPRESENTS THE CORRESPONDING VIEWPOINTS: $0^\circ, 30^\circ, 45^\circ, 60^\circ, 90^\circ$.



From Table X, [4], [27] are 3D-based frontalization solutions that their results would depend on the precision of the facial landmarks detector. [4] did not handle well to fill in the invisible region caused by self-occlusion. Thus, their results may generate ghosting shadows that may influences the performance of expression recognition. The results from [27] are seriously distorted that the facial characteristics, including expression and identity characteristics, are quite different from the ground truth frontal faces. In Table X, the L_2 distance between the synthesized image and the target image is also shown under each synthesized image. It is clear from the table that the proposed model can provide visually appealing face frontalization results, though the L_2 distance to the ground-truth image is not the smallest among all methods under comparison.

Table X
THE FIRST ROW GIVES AN EXAMPLE OF A SMILING FACE UNDER 45° HEAD POSE, AND THE SECOND ROW DEPICTS AN EXAMPLE OF A SCREAMING FACE UNDER 30° HEAD POSE. THE FIRST COLUMN IS THE INPUT IMAGE I^P, THE LAST COLUMN IS THE GROUND TRUTH FRONTAL IMAGE I^F, AND THE MIDDLE COLUMNS ARE THE FACE FRONTALIZATION RESULTS BY USING DIFFERENT FACE FRONTALIZATION APPROACHES. THE L_2 DISTANCE BETWEEN THE SYNTHESIZED IMAGE AND THE TARGET IMAGE IS GIVEN UNDER EACH SYNTHESIZED IMAGE.

Profile I^P	**Ours**	[28]	[4]	[27]	Frontal I^F
					
	711.39	827.21	631.56	920.01	0
					
	594.94	802.04	552.98	886.09	0

V. CONCLUSION

In conclusion, this paper proposed a multi-task GAN-based network model that learns emotion-preserving representation during face frontalization process. The discriminator is trained to distinguish real/fake and recognize class labels. The encoder in the generator learns representative features not only for recognition, but also makes the decoder capable of reconstructing emotion-preserving and realistic-looking frontal faces. By combining several different loss functions, the learnt representations are discriminative for facial expression recognition under large head pose variations, and the synthesized frontal face images maintain the expression characteristics that are effective for recognition task. Experimental results demonstrate that the proposed method outperforms the state-of-the-art facial expression recognition methods on Multi-PIE database.

ACKNOWLEDGMENT

This work was supported by Ministry of Science and Technology, Taiwan, under the project 105-2218-E-007-030.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[4] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4295–4304, 2015.

[5] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *arXiv:1704.04086*, 2017.

[6] Mahdi Jampour, Vincent Lepetit, Thomas Mauthner, and Horst Bischof. Pose-specific non-linear mappings in feature space towards multiview facial expression recognition. *Image and Vision Computing*, 58:38–46, 2017.

[7] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *International Conference on Computer Vision*, pages 2983–2991, 2015.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[9] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv:1609.04802*, 2016.

[10] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, pages 94–101, 2010.

[11] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.

[12] Ali Mollahosseini, David Chan, and Mohammad H Mahoor. Going deeper in facial expression recognition using deep neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1–10, 2016.

[13] S Moore and R Bowden. Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4):541–558, 2011.

[14] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016.

[15] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[16] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015.

[17] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[18] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Lijun Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697, 2012.

[19] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conf. on computer vision and pattern recognition*, pages 1–9, 2015.

[20] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *CVPR*, volume 4, page 7, 2017.

[21] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 676–684, 2015.

[22] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *International Conference on Automatic face and gesture recognition* , pages 211–216. IEEE, 2006.

[23] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[24] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, Jingwei Yan, and Keyu Yan. A deep neural network-driven feature learning method for multi-view facial expression recognition. *IEEE Transactions on Multimedia*, 18(12):2528–2536, 2016.

[25] 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.

[26] Wenming Zheng. Multi-view facial expression recognition based on group sparse reduced-rank regression. *IEEE Transactions on Affective Computing*, 5(1):71–85, 2014.

[27] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 787–796, 2015.

[28] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning identity-preserving face space. In *International Conf. on Computer Vision*, pages 113–120, 2013.