IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 June 2022, accepted 14 July 2022, date of publication 25 July 2022, date of current version 28 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3193386

RESEARCH ARTICLE

Pre-Trained Feature Fusion and Multidomain Identification Generative Adversarial Network for Face Frontalization

SHENGCAI CEN, HAOKUN LUO, JINGHAN HUANG, WURUI SHI, AND XUEYUN CHEN
School of Electrical Engineering, Guangxi University, Nanning 530004, China

Corresponding author: Xueyun Chen (20140043@gxu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 62061002.

ABSTRACT The study of face frontalization is essential for improving face recognition accuracy in extreme pose scenarios. Mainstream methods like TP-GAN, CAPG-GAN, etc., have made meaningful contributions. However, they still suffer from two problems: the lack of extracted feature diversity and the blurred details in generated images. This paper proposes a pre-trained feature fusion and multi-domain identification generative adversarial network (PM-GAN) for face frontalization: the features of the model pre-trained on large-scale datasets are fused with the original features of the encoder to enhance the diversity and robustness of features. In order to fuse features more effectively, we design a novel feature fusion module (FFM). In addition, a group of global and local discriminators is introduced to reinforce the local details and realism of generated frontal faces. Experimental results show that our proposed method outperforms state-of-the-art methods on M^2FPA and CAS-PEAL datasets.

INDEX TERMS Face frontalization, transferring pre-trained network, feature fusion, detail optimization, generative adversarial network.

I. INTRODUCTION

with the development of deep learning, face recognition has made significant progress but still suffers from drastic accuracy reduction in side-view poses. Currently, this problem is mainly tackled in two research directions: the methods of learning pose-invariant features [1]–[3] and the methods of face frontalization [4]–[8]. Since the features learned by the former methods often perform poorly in extreme pose scenarios, face frontalization research has become a hot topic in the face field, and many excellent works have emerged. However, how to generate high-quality faces remains a challenging task.

The methods of face frontalization can be broadly classified into three categories. The first category is based on 2D/3D local texture warping [9]–[11]. Hassner *et al.* [9] used one single and unmodified 3D model for face normalization. The second category is based on statistical modeling [12]–[14]. Sagonas *et al.* [14] used a statistical model to generate frontal faces via tackling a limited low-rank minimization problem. The third category is based on deep neural networks [6], [17], [35]. Kan *et al.* [15] presented progressive stacked autoencoders (SPAE) for gradually converting side-view faces under extreme poses to frontal faces. Zhang *et al.* [16] proposed a flow-based convolutional network for face frontalization to learn the conversion from non-frontal faces to frontal faces in the spatial domain. Benefiting from the excellent generative abilities of generative adversarial networks (GAN), GAN-based face frontalization research is increasing. Hu *et al.* [17] presented a couple-agent pose-guided generative adversarial network (CAPG-GAN) to generate faces of arbitrary angles by using facial landmark heatmaps as pose guidance information. Zhang *et al.* [18] presented a pose-weighted generative adversarial network (PW-GAN) to pay more attention to faces of extreme poses. Duan *et al.* [20] proposed a boosting generative adversarial network(Boost-GAN) to convert side-view faces with external occlusion to frontal faces. Zhang *et al.* [21] introduced an identity-and-pose guided generative adversarial network (IPG-GAN) to enhance identity features. However, most

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

previous generative networks were only trained on face frontalization datasets, which were collected under controlled conditions (limited subjects, poses, lighting, etc.), resulting in insufficient diversity of extracted features. How to extract richer face features is a problem worth studying.

In recent years, many studies [23]–[25] improved the network performance via transferring the models pre-trained on other large-scale and diverse datasets into their architectures. In image synthesis, Johnson *et al.* [26] produced higher-quality images by using a perceptual loss function based on the high-level features of a pre-trained network. Huang *et al.* [6] used the features extracted by a pre-trained face recognition network as identity features to define an identity-preserving loss. The above methods show the effectiveness of using pre-trained models. Most previous works in image synthesis only used the high-level features of pre-trained networks to define loss functions. In contrast, transferring pre-trained models into generative networks to enhance the feature extraction capability of networks is rarely explored and how to better use pre-trained models deserves more in-depth study.

Face synthesis has much higher requirements for local details in critical areas than general image generation. Huang *et al.* [6] proposed a dual-path generative adversarial network (TP-GAN), simultaneously perceiving global structures and local details to develop the detail quality of generated face images. Hao *et al.* [28] proposed a detail-based guidance generative adversarial network for pose reconstruction (DGPR), using dual generators to reconstruct detail and global features. Many researchers improved the detailed quality of synthesis images by designing different structures of generators. Designing a group of local discriminators is another way, but it has not been deeply researched. Yin *et al.* [7] proposed a dual-attention generative adversarial network (DA-GAN), using multiple local discriminators to pay attention to local areas. However, they ignored the differences of importance between different local features. Furthermore, we found that generating clear and consistent facial landmark regions was more difficult than other local features, such as hair and skin color. We consider making the generator pay more attention to those local features that are more difficult to generate.

In response to the above findings, we propose a pre-trained feature fusion and multi-domain identification generative adversarial network (PM-GAN) for face frontalization. To better exploit the capabilities of the pre-trained model and enrich the diversity and robustness of extracted features, we transfer the pre-trained model into the generative network and fuse the features of the pre-trained model with the original features of the encoder in multiple layers. Due to the semantic dissimilarity between pre-trained features and original features, simple fusion operations such as summing and concatenation are challenging to obtain a better performance. So we also design a novel feature fusion module (FFM) to effectively fuse the pre-trained and original features in the channel and spatial dimensions. In addition, we introduce a group of global and local discriminators to encourage the generative network to pay more attention to more important local regions and improve the detail quality of generated images during adversarial training, named multi-domain identification.

Experiments demonstrate that our proposed PM-GAN can generate photorealistic frontal faces with more delicate details and improve the performance of face recognization on M^{2}PFA [29] and CAS-PEAL [30] datasets. Significantly, PM-GAN increases face recognition accuracy by 1.4% and 2.2% under the extreme poses of 75° and 90° on M^{2}FPA, respectively, compared with state-of-the-art methods. Our primary contributions are as follows:

1) We propose a feature optimization method by fusing the features of the pre-trained network with the original features of the encoder to enrich the diversity of extracted features.

2) A novel feature fusion module (FFM) is designed to more effectively fuse features in the channel and spatial dimensions.

3) A image detail enhancement approach based on multi-domain identification is introduced, making the details of generated images clearer.

The rest of this article is organized as follows. We first introduce related works in section 2, and then describe our proposed method in section 3. Finally, experiments and discussions are introduced in section 4 and section 5.

II. RELATED WORK

A. GAN AND FACE FRONTALIZATION

The generative adversarial network (GAN) proposed by Goodfellow *et al.* [31] consists of a generator G and a discriminator D. G takes random noise z as inputs and generates images and D distinguishes generated images $G(z)$ and true images x. GAN can push the distribution of generated images to move toward the distribution of true images during adversarial training. The objective function of GAN can be formulated as follows:

$$\min_G \max_D V(G,D) = E_{x\sim p_{\text{date }(x)}}[\log D(x)] + E_{z\sim p_{\text{date }(z)}}[\log(1-D(G(z))] \quad (1)$$

In recent years, many researchers have made significant progress on GAN, and many excellent GANs emerged, such as pix2pix [32], CGAN [33], Style-GAN [34], etc. Thanks to the generative ability of GAN, GAN-based methods for face frontalization have been the mainstream. Tran *et al.* [35] proposed a feature decoupled learning generative adversarial network (DR-GAN) to learn pose robustness features and generate frontal faces. Inspired by Cycle-GAN [36], Zhang *et al.* [8] presented a cycle-consistent generative adversarial network for face frontalization. A coupled generative adversarial network (PF-cpGAN) was introduced by Taherkhani *et al.* [19] to establish the unseen relationship between side-view faces and frontal faces. Although previous approaches have made significant contributions, there are still problems of insufficient feature diversity and blurred

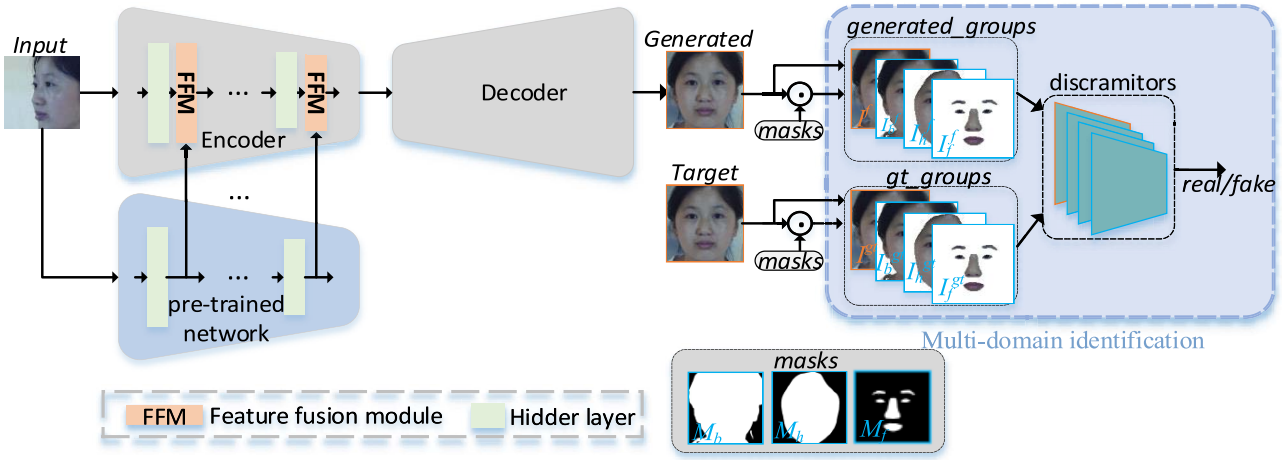

FIGURE 1. Overall framework of our proposed PM-GAN. It contains a generative network and a multi-domain identification. The generative network consists of an encoder-decoder network and a pre-trained network. The features of the pre-trained network are fused with the original features of the encoder by our designed feature fusion modules (FFM) in multiple layers. The multi-domain identification is composed of multiple discriminators capable of discriminating the authenticity of global and local areas of images.

local details in generated images. In this paper, we propose a pre-trained feature fusion and multi-domains identification generative adversarial network (PM-GAN) to increase feature diversity and optimize the local details of images.

B. TRANSFERRING PRE-TRAINED NETWORKS

Transferring networks pre-trained on large-scale datasets to other tasks has been extensively studied. In computer vision, the usage of pre-trained models can be broadly classified into two categories. The first category uses the pre-trained model as an extractor to perform other tasks. Loey *et al.* [42] applied a pre-trained ResNet50 to extract features and send them into another network for performing face mask detection. Iglovikov *et al.* [37] used a pre-trained VGG network as the encoder of U-Net for feature extraction, which improved the robustness of the segmentation model. Dagher *et al.* [22] improved accuracy by using a network pre-trained on face datasets to extract features and send them into an age estimation model. The second category uses the high-level features extracted by the pre-trained network to define feature losses. In the field of face frontalization, Huang [6] and Yin [7] used pre-trained face recognition networks to extract identity features and employed them to define identity-preserving losses. Previous methods indicated the effectiveness of applying pre-trained models to other tasks. However, most existing face frontalization methods applying pre-trained networks mainly used the high-level features of pre-trained networks to define loss functions. How to transfer pre-trained models into generative networks to improve the performance of face frontalization networks is a worthwhile research problem.

To use the pre-trained network to improve the performance of face frontalization, we fuse the features of the network pre-trained on a large-scale face recognization dataset with the original features of the encoder in multiple layers to enrich the diversity of extracted features. Furthermore, considering semantic dissimilarity between pre-trained and original features, we design a novel feature fusion module to fuse features in the spatial and channel dimensions effectively.

III. METHOD

In the section, we first show the overall architecture of our proposed PM-GAN, then describe the pre-trained feature fusion, the designed feature fusion module, and the multi-domain identification. Objective functions, detailed structures of networks, and the algorithm are introduced at the end.

A. NETWORK ARCHITECTURE OF PM-GAN

As shown in Figure 1, the overall network architecture of our proposed PM-GAN can be divided into two parts: a generative network and a multi-domain identification. The generative network is composed of an encoder-decoder and a pre-trained network. The multi-domain identification consists of a group of global and local discriminators.

Side-view faces are fed into the pre-trained network and the encoder. Then, the features extracted by the pre-trained network are fused with the features of the encoder to enhance feature diversity. Finally, the decoder generates frontal faces. Multiple discriminators encourage the generative network to pay more attention to the critical local regions of generated faces during adversarial training.

B. TRANSFERRING PRE-TRAINED MODEL INTO GENERATIVE NETWORK

Most face frontalization networks learn only from face frontalization datasets, which were collected under controlled conditions (limited subjects, poses, lighting, etc.), resulting in the lack of extracted feature diversity. On the contrary, the

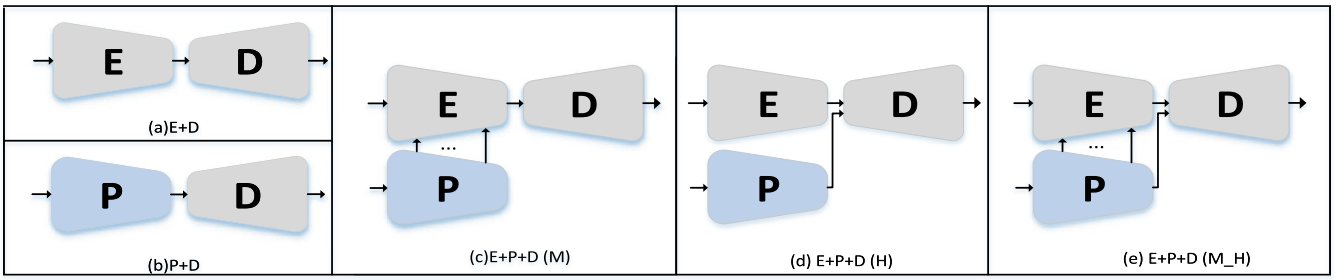

FIGURE 2. Different structures of generative networks. E and D are an encoder and a decoder, respectively. P is a pre-trained model. E+D is a basic encoder-decoder network. P+D means that P replaces E in E+D. (c), (d) and (e) are three various structures that we explore, and they are different in fusion positions. E+D+P (M) is used in PM-GAN.

model pre-trained on large-scale face recognization datasets captured in the wild can extract richer face features. To fully take advantage of the pre-trained model's powerful feature extraction capability, we transfer the pre-trained model into the generative network to extract more diverse features.

Different structures of generative networks are shown in Figure 2. E+D means the encoder-decoder network commonly used in GAN-based methods. P+D uses a pre-trained network for extracting features. Different from the first two structures, E+D+P(M), E+D+P(H), and E+D+P(H_M) use P and E to extract features in parallel. Especially, E+D+P(M) uses the pre-trained network and encoder to extract features and fuses them in middle layers, which we finally use in PM-GAN. E+D+P(H) fuses features in high-level layers while E+D+P(M_H) in middle and high-level layers.

C. FEATURE FUSION MODULE

Feature fusion is commonly performed via simple operations such as summing or concatenation. However, the semantic dissimilarity between different features is ignored, resulting in poor fusion performance. Considering the semantic differences between the pre-trained features and original features, we design a novel feature fusion module to smooth the semantic dissimilarity between the two features in the channel and spatial dimensions, fusing pre-trained features with original features more efficiently.

Our proposed feature fusion module (FFM) is shown in Figure 3. FFM-a and FFM-b are two different feature fusion modules proposed in the paper. The channel modulation module (CMM) and spatial modulation module (SMM) are the components of FFM.

$f_o \in R^{C\times H\times W}$ and $f_p \in R^{C_1\times H\times W}$ denote the original features and pre-trained features, respectively. $f_{o+1} \in R^{2C\times H\times W}$ is the output of FFM. CMM and SMM aim to obtain the channel modulation map M_c and spatial modulation map M_s. M_c and M_s are defined as follows:

$$M_c = \Phi_{\sigma c}\left(f_{po1}\right) \tag{2}$$

$$M_s = \Phi_{\sigma s}\left(f_{po2}\right) \tag{3}$$

$$f_{o+1} = \Phi_c\left(f_o \otimes (M_c + 1) \otimes (M_s + 1),\right.$$
$$\left. f_p \otimes (M_c + 1) \otimes (M_s + 1)\right) \tag{4}$$

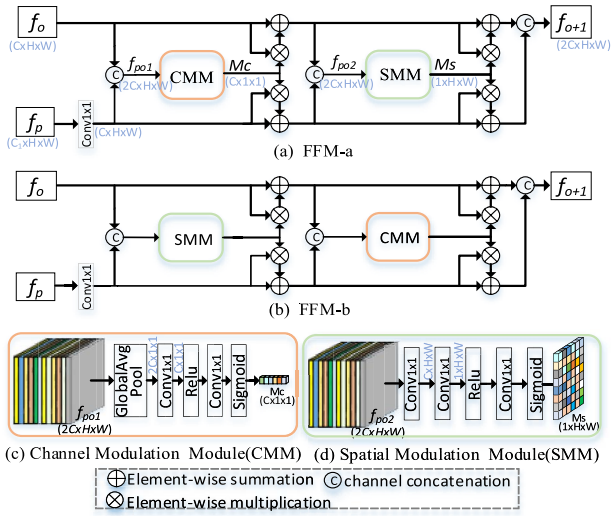

FIGURE 3. Feature Fusion Module (FFM). FFM-a and FFM-b are two various feature fusion modules proposed in the paper. CMM and SMM are the channel modulation module and spatial modulation module separately, and they are the components of FFM-a and FFM-b. FFM-a is used in PM-GAN.

where f_{po1} and f_{po2} denote the simple fusion of pre-trained features and original features through channel concatenation. $\Phi_{\sigma c}(\cdot)$ and $\Phi_{\sigma s}(\cdot)$ are defined as the calculation processes of M_c and M_s, respectively. FFM-a can be formulated as equation 4. $\Phi_c(\cdot)$ is channel concatenation and $\otimes$ is element-wise multiplication.

D. MULTI-DOMAIN IDENTIFICATION

Face image is classified into four components: facial landmark regions, skin color, hair, and background. In this paper, we consider the importance difference of different local features and pay more attention to more important local features. Through experimental studies, we found that it is more difficult to generate clear and consistent facial landmark regions than skin color and hair, so we consider facial landmark regions as the most important face component and pay more attention to them.

Different structures of identification networks are shown in Figure 4. 1D only uses a global discriminator to discriminate the authenticity of global face images. 2D based on 1D adds a local discriminator to distinguish the authenticity of face

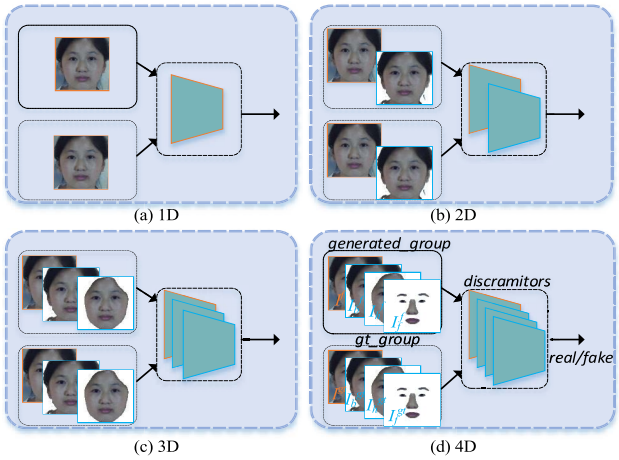

FIGURE 4. Different structures of identification networks. 1D, 2D, 3D, and 4D use different numbers of discriminators for identifying images. 1D is commonly used in GAN-based methods. 2D, 3D, and 4D are various structures we explore. Specifically, 4D is used in PM-GAN.

images without the background. Based on 2D, 3D applies another discriminator for discriminating the authenticity of local face images without background or hair. We add a facial landmarks regions discriminator on top of 3D to form 4D, which is used in PM-GAN.

4D shows that all global and local images contain the facial landmark regions. Four discriminators push the generative network to pay more attention to the facial landmark regions, while the background is only concerned by the global discriminator during adversarial training.

To obtain local images, we feed target images I^{gt} into a face attribute segmentation network f_p and obtain multiple local masks, including background-free mask M_b, mask without hair or background M_h, and facial landmarks regions mask M_f, which is defined as equation 5.

$$M_b, M_h, M_f = f_p\left(I^{gt}\right) \tag{5}$$

$$I_b^{gt}, I_h^{gt}, I_f^{gt} = \left(M_b, M_h, M_f\right) \odot I^{gt} \tag{6}$$

$$I_b^{f}, I_h^{f}, I_f^{f} = \left(M_b, M_h, M_f\right) \odot I^{f} \tag{7}$$

Then target images I^{gt} and generated images I^f are processed with the three masks, respectively. Finally, we obtain corresponding target local images $(I_b^{gt}, I_h^{gt}, I_f^{gt})$ and generated local images (I_b^f, I_h^f, I_f^f). The processing can be formulated as equation 6 and equation 7. $\odot$ means the element-wise product. The global and local images are fed into corresponding discriminators for identification.

E. OBJECTIVE FUNCTIONS

PM-GAN learns a series of objective functions in a supervised learning manner. The overall objective function of the generative network is composed of multiple losses: a multi-scale pixel-wise loss, an adversarial loss, a total variation regularization loss, an identity preserving loss, and a symmetry loss. The discriminators are only supervised by the adversarial loss.

1) MULTI-SCALE PIXEL-WISE LOSS

To keep the content consistency between generated images I^f and target images I^{gt}, we use a multi-scale pixel-wise loss to supervise the generative network. It can be defined as follow:

$$L_{pix} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{C \times W_s \times H_s}\sum_{c,h_s,w_s=1}^{C,H_s,W_s}\left|I^f_{s,c,h_s,w_s} - I^{gt}_{s,c,h_s,w_s}\right| \tag{8}$$

where $|\cdot|$ denotes the L1-norm. S means the number of scales, and W_s and H_s are the corresponding width and height at the scale s and the channel c. S is set as 3, and the scales are 32×32, 64×64, and 128×128, respectively.

2) ADVERSARIAL LOSS

The pixel-wise loss produces overly smooth images, while the adversarial loss can help to generate more realistic images and make the distribution of generated images move close to the distribution of true images. The overall adversarial loss of the generative network G and discriminators D can be expressed as follow:

$$\min_G \max_D V(G,D) = \sum_{j\in\{g,b,h,f\}}\left(E_{I_j^{gt}}\left[\log D_j\left(I_j^{gt}\right)\right] + E_{I_j^f}\left[1 - \log D_j\left(I_j^f\right)\right]\right) \tag{9}$$

where I_g^f and I_g^{gt} represent generated images and target images, $\{I_b^f, I_h^f, I_f^f\}$ and $\{I_b^{gt}, I_h^{gt}, I_f^{gt}\}$ mean the local images. D_g is the global discriminator while $\{D_b, D_h, D_f\}$ are local discriminators.

$$L_{advg} = -\sum_{j\in\{g,b,h,f\}} E_{I_j^f}\left[\log D_j\left(I_j^f\right)\right] \tag{10}$$

$$L_{advd} = -\sum_{j\in\{g,b,h,f\}}\left(E_{I_j^{gt}}\left[\log D_j\left(I_j^{gt}\right)\right] + E_{I_j^f}\left[1 - \log D_j\left(I_j^f\right)\right]\right) \tag{11}$$

The adversarial loss can be formulated as equation 10 when optimizing the generative network G. Similarly, the objective function of the discriminators D can be formulated as equation 11.

3) TOTAL VARIATION REGULARIZATION LOSS

We use a total variation regularization loss to eliminate undesirable artifacts in generated images. It is defined as follows:

$$L_{tv} = \frac{1}{C \times H \times W}\sum_{c,h,w=1}^{C,H-1,W-1}\left(\left|I^f_{c,h,w} - I^f_{c,h,w+1}\right| + \left|I^f_{c,h,w} - I^f_{c,h+1,w}\right|\right) \tag{12}$$

where $I^f_{c,h,w}$ and $I^f_{c,h,w+1}$ denote adjacent pixels in the width dimension. Similarly, $I^f_{c,h,w}$ and $I^f_{c,h+1,w}$ are adjacent pixels in the height dimension.

4) IDENTITY PRESERVING LOSS

In order to preserve more identity information, our generative network is also supervised by an identity preserving loss, which can be formulated as follows:

$$L_{id} = \frac{1}{2}\sum_{i=1}^{2}\left\|f_{ID}\left(I^{f}\right) - f_{ID}\left(I^{gt}\right)\right\| \tag{13}$$

where $\|\cdot\|$ is the L2-norm and $f_{ID}(\cdot)$ means the output features of the last two layers of the pre-trained face recognition LightCNN [38]. I^{f} and I^{gt} are generated images and target images, respectively.

5) SYMMETRY LOSS

The symmetry loss proposed by TP-GAN can remove the unnatural asymmetry effect of generated faces. It can be defined as follow:

$$L_{sym} = \frac{2}{H \times W}\sum_{h,w=1}^{H,W/2}\left(\left|I^{f}_{h,w} - I^{f}_{h,W-w}\right|\right) \tag{14}$$

where $I^{f}_{h,w}$ and $I^{f}_{h,W-w}$ denote the left-right symmetric pixels in the width dimension.

6) OVERALL OBJECTIVE FUNCTION OF GENERATIVE NETWORK

The total loss of the generative network is a weighted sum of the losses mentioned above. It can be formulated as follows:

$$L_G = \lambda_1 L_{pix} + \lambda_2 L_{advg} + \lambda_3 L_{id} + \lambda_4 L_{sym} + \lambda_5 L_{tv} \tag{15}$$

where λ_1, λ_2, λ_3, λ_4, and λ_5 are the hyperparameters that regulate the weight of different losses.

F. THE DETAILED STRUCTURES OF NETWORKS

The detailed structures of the encoder-decoder and discriminator in PM-GAN are listed in Table 1. Each convolution (Conv) and transposed convolution (TConv) are followed by batch normalization (BN) and rectified linear unit (ReLU) except TConv42. In particular, BN and hyperbolic tangent unit (Tanh) are used after TConv42. The sigmoid activate function is employed after the full connection layer (FC).

TABLE 1. Structures of the encoder-decoder and the discriminator. The $layer^{*E/D}$ only is used in the encoder/ discriminator.

Encode/Discriminator		Decoder	
Layer (Filter/Stride)	Output Size (CxHxW)	Layer (Filter/Stride)	Output Size (CxHxW)
Conv11(3x3/1)	32x128x128	TConv11(3x3/2)	512x32x32
FFM_1$(-)^{*E}$	64x128x128	-	-
Conv12(3x3/1)	64x128x128	TConv12(3x3/1)	256x32x32
Conv21(3x3/2)	128x64x64	TConv21(3x3/2)	128x64x64
FFM_2$(-)^{*E}$	256x64x64	-	-
Conv23(3x3/1)	128x64x64	TConv22(3x3/1)	128x64x64
Conv31(3x3/2)	256x32x32	TConv31(3x3/2)	64x128x128
FFM_3$(-)^{*E}$	512x32x32	-	-
Conv32(3x3/1)	256x32x32	TConv32(3x3/1)	64x128x128
Conv41(3x3/2)	512x16x16	TConv41(3x3/1)	32x128x128
FFM_4$(-)^{*E}$	1024x16x16	-	-
Conv42(3x3/1)	512x16x16	TConv42(3x3/1)	3x128x128
AvgPool^{*D}	512x1x1	-	-
FC$(-)^{*D}$	1	-	-

Algorithm 1 Pre-Trained Feature Fusion and Multi-Domain Identification Generative Adversarial Network for Face Frontalization

Input: Side-view faces I^{p}, true frontal faces I^{gt}. θ_G and θ_D are the parameters of the generative network G and discriminators D, respectively.

Output: Generated frontal faces I^{f}

for the number of iterations **do**
- Randomly select m pairs date $\{I^{p}, I^{gt}\}$;
- Feed I^{p} into G, and G generates frontal faces I^{f};
- Obtain multiple local images $\{I^{f}_{b}, I^{f}_{h}, I^{f}_{f}\}$ after masks processing;
- Feed $\{I^{f}, I^{f}_{b}, I^{f}_{h}, I^{f}_{f}\}$ into D;
- Calculate the cost of the function L_D of D according to equation 11;
- Update the parameters θ_D by using the Adam optimization algorithm: $\theta_D = \text{Adam}\left(\nabla_{\theta_D}(L_D), \theta_D\right)$
- Calculate the cost of the function L_G of G according to equation 15;
- Update the parameters θ_G by using the Adam optimization algorithm: $\theta_G = \text{Adam}\left(\nabla_{\theta_G}(L_G), \theta_G\right)$

end for

G. ALGORITHM

The specific algorithm flow for training our proposed PM-GAN is shown in Algorithm 1.

IV. EXPERIMENTS

To illustrate that PM-GAN can synthesize photorealistic images with finer detail while preserving identity. We first introduce datasets, implementation details, and evaluation metrics in the following subsections. Then, we conduct ablation studies to prove the advantages of the three components proposed in the paper and design experiments compared with current state-of-the-art methods. Finally, we qualitatively evaluate the equality of generated images.

A. EXPERIMENTAL SETTINGS

1) DATASETS

We conducted our experiments on M^{2}FPA [29] and CAS-PEAL [30] datasets. M^{2}FPA was released in 2019 and is a public dataset for facial pose research. It has 397,544 images containing 229 subjects with 13 yaws ($-90^\circ \sim +90^\circ$), five pitches ($-30^\circ \sim +45^\circ$), four attributes, and seven illumination changes. We only used the images under 13 yaws of 0° pitch for the following experiments. There are face frontalization benchmarks on M^{2}FPA, including TP-GAN [6], DR-GAN [35], and CAPG-GAN [17]. We follow the official settings for our experiments.

CAS-PEAL [30] is a public dataset available for facial pose research. It is composed of 30,863 grayscale images of 1040 subjects with nine poses. Following previous methods, we used the pictures with angles of 0°, ±15°, ±30°, and ±45° for a fair comparison. It is divided into train/test sets with 7 : 3 at random.

2) IMPLEMENTATION DETAILS

Following previous methods, all images were cropped and resized to 128 × 128. Image intensities were scaled to the range of [−1, 1]. In our experiments, we set $\lambda_1 = 10$, $\lambda_2 = 0.1$, $\lambda_3 = 0.3$, $\lambda_4 = 0.1$, $\lambda_5 = 0.01$. We used the Adam optimizer with the β_1 of 0.5 and β_2 of 0.99. The learning rate was initialized with 2e-4 and gradually decreased after each epoch until it reached 0. The batch size was set as 32, and all trainable parameters of networks were initialized by a normal distribution. We implemented our code with PyTorch and trained our PM-GAN on four GeForce GTX 3090 GPUs.

3) EVALUATION METRICS

Following previous works, the rank-1 recognition rate is mainly used to evaluate the face recognition performance of the proposed method. We generate frontal faces and send them to a pre-trained face recognition network LightCNN [38] for extracting deep features. The rank-1 recognition rate is calculated with the cosine distance of the deep features. In addition, We also use peak signal-to-noise ratio (PSNR) to evaluate the quality of generated images and structural similarity index (SSIM) for measuring structural similarity between generated images and target images.

B. ABLATION STUDIES

1) THE EFFECTS OF TRANSFERRING PRE-TRAINED MODEL

To clarify the effectiveness of transferring the pre-trained model into the generative network on face frontalization. As shown in Table 2, we conducted experiments on different structures of generative networks. We used a FaceNet [39] pre-trained on VGGFace2 [40] dataset as the pre-trained model P. The discriminator was the standard 1D in Figure 4, and the fusion operation used the typical concatenation. We explored different structures of generative networks.

We find that transferring the pre-trained model into the generative network can significantly improve the rank-1 recognition rate performance of face frontalization. It is worth noting that E+D+P(M) works better than other network structures. Specifically, It increases the rank-1 recognition rate by 8.3% and 8% under angles of ±75° and ±90°, separately, compared with the baseline E+D. Experimental results illustrate that fusing the features of the pre-trained model with the original features of the encoder can improve the performance of face frontalization, and feature fusion works best in the middle layers. We think that transferring the pre-trained model to the generative network improves the feature extraction capability of the network and enhances the diversity of the extracted features. We use E+D+P (M) as the generative network structure of PM-GAN.

TABLE 2. Rank-1 recognition rate (%) performance of different structures of generative networks in Figure 2 on M^{2}FPA.

Network structure	±15°	±30°	±45°	±60°	±75°	±90°
E+D	99.2	96.6	92.3	90.4	80.2	60.4
P+D	99.3	97.8	96.1	93.4	83.2	63.2
E+D+P(M)	**99.6**	**98.9**	**98.4**	**97.2**	**88.5**	**68.4**
E+D+P(H)	99.5	97.9	97.2	96.8	86.8	66.5
E+D+P(M_H)	99.6	98.6	98.2	97.1	87.7	67.9

TABLE 3. Rank-1 recognition rate (%) performance of different pre-trained models on M^{2}FPA.

Pre-trained model	±15°	±30°	±45°	±60°	±75°	±90°
LightCNN	99.5	98.7	98.3	97.0	88.2	68.3
SphereFace	99.4	98.5	98.0	96.9	88.1	67.8
FaceNet	**99.6**	**98.9**	**98.4**	**97.2**	**88.5**	**68.4**

TABLE 4. Rank-1 recognition rate (%) performance of different fusion operations on M^{2}FPA. FFM-a and FFM-b mean the structure graphs in Figure 3.

Fusion operation	±15°	±30°	±45°	±60°	±75°	±90°
Concatenation	99.6	98.9	98.4	97.2	88.5	68.4
Summing	99.4	98.5	97.9	97.0	88.4	67.8
FFM-a	**99.9**	**99.4**	**99.2**	**98.4**	**88.6**	**68.8**
FFM-b	99.8	99.3	99.2	98.3	88.6	68.7

To further investigate the effect of different face recognition models as pre-trained networks on face frontalization, we compared three different pre-trained models: LightCNN [38], SphereFace [41], and FaceNet [39]. Experimental results are reported in Table 3. It can be seen that the rank-1 recognition rate of different pre-trained models is a small gap, but FaceNet works slightly better than LightCNN and SphereFace. FaceNet is selected as the pre-trained network of PM-GAN.

2) THE EFFECTS OF FEATURE FUSION MODULE

To prove the effectiveness of our proposed feature fusion module (FFM), we designed experiments shown in Table 4. We choose the E+D+P(M) in Table 2 as the basic generative network structure, but they fuse features with different fusion operations. Compared with typical fusion operations (Summing and Concatenation), our proposed FFM-a and FFM-b can improve the rank-1 performance rate across all poses on M^{2}FPA. In particular, FFM-a improves the rank-1 recognition rate by 0.8% and 1.2% under angels of ±45° and ±60°, respectively, compared with the concatenation fusion operation. Experimental results demonstrate that FFM can effectively fuse pre-trained features and original features and improve the performance of face frontalization. We think that it is because the proposed feature fusion module can better smooth the semantic differences of different features and obtain a better fusion performance. Finally, we choose FFM-a as the fusion operation in PM-GAN.

TABLE 5. Rank-1 recognition rate (%) performance of different identification networks in Figure 4 on M^2FPA.

nD	±15°	±30°	±45°	±60°	±75°	±90°
1D	99.9	99.4	99.2	98.4	88.6	68.8
2D	99.9	99.6	99.5	98.6	88.9	69.5
3D	100	99.9	99.7	98.8	89.3	70.1
4D	**100**	**100**	**99.8**	**98.8**	**89.5**	**70.3**

TABLE 6. PSNR and SSIM performance of different identification networks in Figure 4 on M^2FPA.

nD	PSNR	SSIM
1D	21.88	0.69
2D	22.35	0.71
3D	22.90	0.75
4D	**23.18**	**0.76**

3) THE EFFECTS OF MULTI-DOMAIN IDENTIFICATION

To verify the contributions of our proposed multi-domain identification to face frontalization, we designed experiments shown in Table 5. We use E+D+P(M) as the generative network, and the feature fusion operation is FFM-a. Different structures of identification networks are explored. Experimental results show that increasing the count of discriminators from one to four can gradually improve the face recognition accuracy across all poses on M^2FPA. Noteworthy, four discriminators (4D) achieves the best performance. Compared to 1D commonly used in GAN-based methods, our proposed 4D obtains accuracy improvements of 0.9% and 1.5% under angles of ±75° and ±90°. It is because the group of global and local discriminators makes the generative network pay more attention to the important local regions in the face image. The above analysis illustrates the validity of the multi-domain identification.

As shown in Table 6, it can be seen that 4D increases the PSNR score and SSIM score by 1.3 and 0.07 on M^2FPA, compared with 1D. The results demonstrate that our proposed multi-domain identification can generate higher-quality images and constrain face structure consistency.

4) THE VISUAL RESULTS OF GRADUALLY ADDING PROPOSED COMPONENTS

To visually show the impact of transferring the pre-trained model, the feature fusion module, and the multi-domain identification on face frontalization, we gradually add the pre-trained network (FaceNet), FFM-a, and the multi-domain identification to the basic network (G_1+1D), which only contains an encoder-decoder (G_1) and a discriminator (1D). Generated results are shown in Figure 5. We observe that G_1+P+1D can generate faces with better identity consistency than G_1+1D. After adding FFM, G_1+P+FFM+1D can generate face images with more significant identity features than G_1+P+1D, but generated images are blurry in some areas. After adding 4D, G_1+P+FFM+4D can generate face images with clearer details than G_1+P+FFM+1D. Experimental results qualitatively illustrate the benefits of each component proposed in this paper.

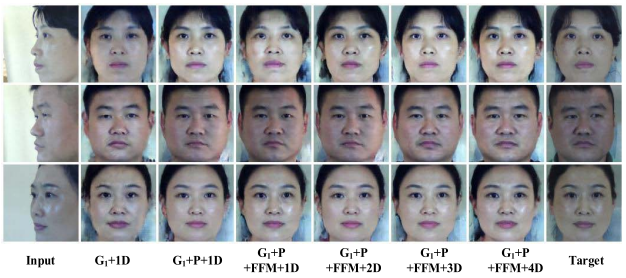

FIGURE 5. Generated results of gradually adding the pre-trained network (P), FFM, and the multi-domain identification to a basic network (G_1+1D). G_1 means E+D. G_1+P represents E+D+P(M). Specifically, G_1+P+FFM+4D is our PM-GAN.

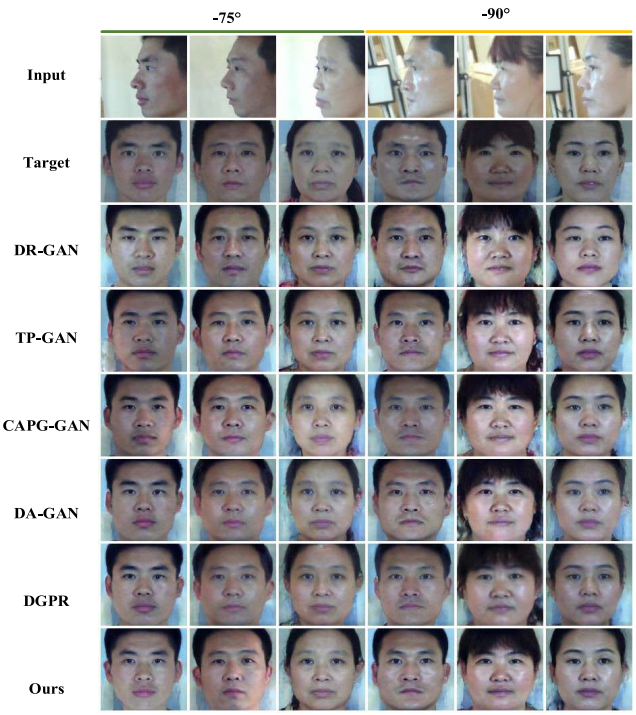

FIGURE 6. Generated results of different models under extreme poses on M^2FPA.

TABLE 7. Rank-1 recognition rate (%) performance of compared models on M^2FPA.

Model	±15°	±30°	±45°	±60°	±75°	±90°
Original	100	100	99.8	98.6	86.9	51.7
DR-GAN	98.9	97.9	95.7	89.5	70.3	35.5
TP-GAN	99.9	99.8	99.1	97.3	87.6	62.1
CAPG-GAN	99.9	99.7	99.4	96.4	87.2	63.9
DA-GAN	99.9	99.9	99.5	97.8	87.7	67.8
DGPR	100	100	99.6	98.6	88.1	68.1
PM-GAN	**100**	**100**	**99.8**	**98.8**	**89.5**	**70.3**

C. COMPARISON EXPERIMENTS

To further prove the identity preservation ability of our proposed method. Table 7 lists the rank-1 recognition rate of our method compared with existing state-of-the-art methods on M^2FPA. The rank-1 recognition rate gradually decreases as angles increase from 0° to 60°. It sharply drops while

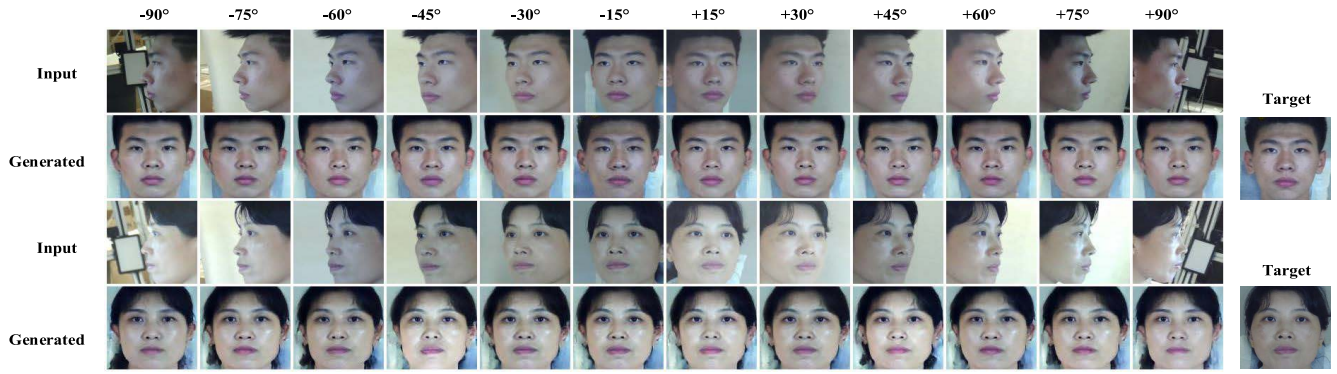

FIGURE 7. Synthesized results of PM-GAN across poses $(-90° \sim +90°)$ on M^{2}FPA.

TABLE 8. Rank-1 recognition rate (%) performance of compared models on CAS-PEAL.

Model	±15°	±30°	±45°	Avg
CR-GAN	97.61	95.80	89.73	94.38
TP-GAN	100	99.94	98.71	99.55
DA-GAN	100	100	99.70	99.90
DGPR	100	100	99.84	99.95
PM-GAN	**100**	**100**	**99.92**	**99.97**

angles are larger than 60°. It is because side-view faces under larger angles provide less helpful information. Notably, our proposed method achieves higher rank-1 recognition accuracy than state-of-the-art methods across most poses. In particular, our PM-GAN raises the rank-1 recognition accuracy by 1.4% and 2.2% under extreme angles of ±75° and ±90°, respectively. We think that it is because our model can extract richer face features and generate clearer and more consistent local details.

Table 8 reports the rank-1 recognition rate of PM-GAN on CAS-PEAL, compared with the state-of-the-art methods, including CR-GAN [27], TP-GAN [6], DA-GAN [7], and DGPR [28]. The comparison results demonstrate the superior performance of our PM-GAN.

To visually report the performance of PM-GAN, Figure 6 shows the generated results of our proposed method and some state-of-the-art methods on M^{2}FPA. It can be seen that all of the above methods can generate realistic images, but it is remarkable that our proposed PM-GAN can generate frontal faces with better identity preservation and clearer details than other models. It further illustrates that our PM-GAN can extract richer features and generate higher quality faces.

D. FACE SYNTHESIS

To qualitatively demonstrate the generative ability of our proposed PM-GAN. Figure 7 and Figure 8 show synthesized results of the persons under arbitrary poses on M^{2}FPA and CAS-PEAL datasets separately. It can be seen that our proposed method can generate photorealistic frontal faces with high identity consistency under the angles within 60°. When the angle is larger than 60°, although the similarity between view-side faces and frontal faces decreases severely,

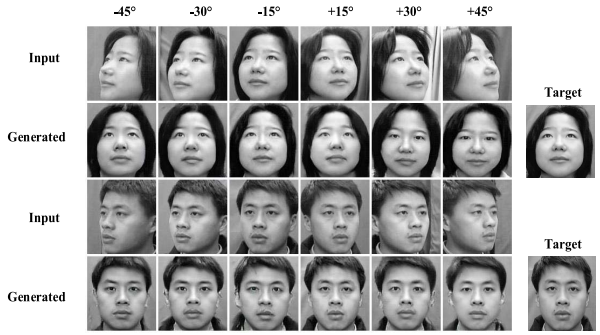

FIGURE 8. Synthesized results of PM-GAN across poses $(-45° \sim +45°)$ on CAS-PEAL.

FIGURE 9. Failure cases. Failure results in the case of the extreme pose and intense illumination.

facial landmark regions and structures of the generated images still maintain high similarity with target images. The visual perception proves that our PM-GAN can extract more identity robust features and generate photorealistic frontal faces with identity preservation under extreme poses.

Although our method can perform well in most cases, some challenges still exist under intense illumination and extreme poses scenarios. Figure 9 shows some failure cases. Generated faces perform poorly in some local areas since the facial information of input images is seriously missing, such as the features of the nose and mouth are covered by intense illumination. However, it can still recover most of the

facial features accurately. It further proves the superiority of PM-GAN in identity preservation and image quality.

V. CONCLUSION

In this paper, aiming at the two weaknesses of the lack of extracted feature diversity and blurred details in generated images, we present pre-trained feature fusion and multi-domain identification generative adversarial network (PM-GAN) for face frontalization. The features of the network pre-trained on large-scale face recognition datasets are fused with the encoder's original features to enhance the features' diversity and robustness. We also design a novel feature fusion module (FFM) to make the feature fusion more effective. A method based on multi-domain identification for optimizing details is proposed to improve the detail quality of generated images. Experiments demonstrate that our proposed PM-GAN can synthesize photorealistic frontal faces with finer detail and improve face recognition performance, especially in extreme pose scenarios. In future work, we will conduct further research on face frontalization in more complex environments with intense light and extreme poses.

REFERENCES

[1] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2049–2058, Nov. 2015.

[2] Q. Meng, X. Xu, X. Wang, Y. Qian, Y. Qin, Z. Wang, C. Zhao, F. Zhou, and Z. Lei, "PoseFace: Pose-invariant features and pose-adaptive loss for face recognition," 2021, *arXiv:2107.11721*.

[3] C. Ding and D. Tao, "Pose-invariant face recognition with homography-based normalization," *Pattern Recog.*, vol. 66, pp. 144–152, Jun. 2017.

[4] X. Luan, H. Geng, L. Liu, W. Li, Y. Zhao, and M. Ren, "Geometry structure preserving based GAN for multi-pose face frontalization and recognition," *IEEE Access*, vol. 8, pp. 104676–104687, 2020.

[5] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 676–684.

[6] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2458–2467.

[7] Y. Yin, S. Jiang, J. P. Robinson, and Y. Fu, "Dual-attention GAN for large-pose face frontalization," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 249–256.

[8] Z. Zhang, R. Liang, X. Chen, X. Xu, G. Hu, W. Zuo, and E. R. Hancock, "Semi-supervised face frontalization in the wild," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 909–922, 2021.

[9] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4295–4304.

[10] L. A. Jeni and J. F. Cohn, "Person-independent 3D gaze estimation using face frontalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 792–800.

[11] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4010–4019.

[12] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3D morphable models," *Int. J. Comput. Vis.*, vol. 126, pp. 233–254, Apr. 2018.

[13] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. J. Dunaway, "A 3D morphable model learnt from 10,000 faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5543–5552.

[14] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3871–3879.

[15] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive auto-encoders (SPAE) for face recognition across poses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1883–1890.

[16] Z. Zhang, X. Chen, B. Wang, G. Hu, W. Zuo, and E. R. Hancock, "Face frontalization using an appearance-flow-based convolutional neural network," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2187–2199, May 2019.

[17] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8398–8406.

[18] S. Zhang, Q. Miao, M. Huang, X. Zhu, Y. Chen, Z. Lei, and J. Wang, "Pose-weighted GAN for photorealistic face frontalization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2384–2388.

[19] F. Taherkhani, V. Talreja, J. Dawson, M. C. Valenti, and N. M. Nasrabadi, "PF-cpGAN: Profile to frontal coupled GAN for face recognition in the wild," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–10.

[20] Q. Duan and L. Zhang, "Look more into occlusion: Realistic face frontalization and recognition with BoostGAN," *IEEE Trans. Netw. Learn. Syst.*, vol. 32, no. 1, pp. 214–228, Jan. 2021.

[21] Y. Zhang, K. Fu, C. Han, and P. Cheng, "Identity-and-pose-guided generative adversarial network for face rotation," *Neurocomputing*, vol. 450, pp. 33–47, Aug. 2021.

[22] I. Dagher and D. Barbara, "Facial age estimation using pre-trained CNN and transfer learning," *Multimedia Tools Appl.*, vol. 80, no. 13, pp. 20369–20380, May 2021.

[23] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Feature transfer learning for face recognition with under-represented data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5697–5706.

[24] M. Long, H. Zhu, J. Wang, and M. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 136–144.

[25] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.

[26] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 694–711.

[27] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, S. Yan, and J. Feng, "Towards pose invariant face recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 221–2207.

[28] J. Hao and X. Chen, "Detailed feature guided generative adversarial pose reconstruction network," *IEEE Access*, vol. 9, pp. 56093–56103, 2021.

[29] P. Li, X. Wu, Y. Hu, R. He, and Z. Sun, "M2FPA: A multi-yaw multi-pitch high-quality dataset and benchmark for facial pose analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10042–10050.

[30] W. Gao, B. Cao, S. Shan, X. Chen, D. X. Zhou Zhang, and D. Zhao, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 1, pp. 149–161, Jan. 2008.

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[32] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[33] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[34] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

[35] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1283–1292.

[36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[37] V. Iglovikov and A. Shvets, "TernausNet: U-Net with VGG11 encoder pre-trained on ImageNet for image segmentation," 2018, *arXiv:1801.05746*.

[38] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[39] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[40] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.
[41] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6738–6746.
[42] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic," *Measurement*, vol. 167, Jan. 2021, Art. no. 108288.

SHENGCAI CEN received the B.E. degree from Guangxi University, China, in 2020, where he is currently pursuing the master's degree with the College of Electrical Engineering. His current research interests include face recognition and computer vision.

HAOKUN LUO received the B.E. degree in automation from the Guilin University of Electronic Technology, China, in 2018. Currently, he is a Graduate Student with the School of Electrical Engineering, Guangxi University. During the period in Guangxi University, his main research interest includes face attribute modification, especially the problem of face posture correction based on generative adversarial networks.

JINGHAN HUANG received the B.E. degree in automation from Guangxi University, Nanning, China, in 2019, where he is currently pursuing the M.S. degree in electronic information. His research interests include computer game, image processing, and deep reinforcement learning.

WURUI SHI received the B.E. degree from Nanjing Agricultural University, Nanjing, China, in 2020. He is currently pursuing the M.S. degree with Guangxi University. His research interests include object detection, computer vision, and deep learning.

XUEYUN CHEN is currently an Associate Professor and the Ph.D. Supervisor with the School of Electrical Engineering, Guangxi University. His research interests include target detection and recognition in remote sensing image, face detection, road detection, and automatic driving.