Face Frontalization with Deep Gan via Multi-Attention Mechanism

Jiaqian Cao
Shandong University
Zhenxue Chen (chenzhenxue@sdu.edu.cn)
Shandong University
Yujiao Zhang
Shandong University
Luna Sun
Shandong University
Jiyang Chen
Shandong University

Research Article

Keywords: Face Frontalization, Generative Adversarial Network, Deep Feature Encoder, Attention Mechanism

Posted Date: August 5th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1888236/v1

Face Frontalization with Deep GAN via Multi-attention Mechanism

Jiaqian Cao[1], Zhenxue Chen[1†], Yujiao Zhang[1], Luna Sun[1], Jiyang Chen[2]

(1. School of Control Science and Engineering, Shandong University, Jinan 250061, China;

2. Institute of Marine Science and Technology, Shandong University, Qingdao, Shandong, 266237, China)

† Corresponding author: chenzhenxue@sdu.edu.cn

Abstract—In recent years, the development of deep learning has led to some advances in face synthesis approaches, but significant pose is remains one of the factors that is difficult to overcome. Benefiting from the proposal and development of generative adversarial network, the level of face frontalization technology has reached new heights. In this paper, we propose a deep generative adversarial network based on multi-attention mechanism (DMA-GAN) for multi-pose face frontalization. Specifically, we add a deep feature encoder based on the attention mechanism and residual block in the generator, which can deepen the network to extract more detailed features and make full use of the long-range dependencies between local features to generate better identity-preserving faces. Meanwhile, to carry the global and local facial information, the discriminator of our model consists of four independent discriminators. The self-attention mechanism is also added to these discriminators to provide more accurate synthesized details. The results from quantitative and qualitative experiments on CAS-PEAL-R1 dataset show that our model proves effective.

Index Terms—Face Frontalization, Generative Adversarial Network, Deep Feature Encoder, Attention Mechanism

I. INTRODUCTION

With the rapid development of deep learning, the performance of face recognition techniques has improved significantly [1] [2]. However, in practical applications (such as surveillance video), face images are often affected by multiple poses. The pose variation in face images greatly reduces the accuracy of most face recognition algorithms, especially in extreme angles, pitches and yaws situations.

At present, multi-pose face recognition methods for multi-pose are divided into two categories. One category entails learning pose-invariant features from original face images [3] [4] [5]. Due to the deletion of features in large pose cases and the variety of postures that may be included, this category has some limitations. The other category involves synthesizing an identity-preserved frontal-view face image from a face image in a specific pose, which is called face frontalization. Then it uses the generated face images to extract features and recognize the face. Previous works [6] [7] [8] all show strong face recognition performance using this method. Since Goodfellow et al. [9] proposed the generative adversarial networks,

GAN has been widely used in the field of image generation and has greatly improved the quality of generated images. In recent years, many face frontalization methods based on GAN have been proposed [10] [11] [12] [13] [14] [15] [16] [17] [18]. The generator of GAN extracts the features of the non-normal face (source) and synthesizes the normal face (target). The discriminator of GAN authenticates the synthetic face images and encourages the generator to synthesize more realistic images. In recent years, many studies have proven that the attention mechanism can improve the performance of the network. Attention mechanisms have been widely used in the field of image processing and achieve satisfactory performance [19] [20] [21] [22]. Inspired by these methods, we propose a deep generative adversarial network based on multi-attention mechanism (DMA-GAN). The visualization of generated results by our model is shown in Figure 1.

In our model, the generation network is based on U-net architecture and is combined with a deep feature encoder (DFE). The deep feature encoder model is grounded on four residual attention modules, and each residual attention module is composed of a residual block and dual-attention module (Fig. 2). Using the residual block can deepen the network to fully extract the deep features on the premise of maintaining good model performance. The DFE can help synthesize frontal face images that better preserve more geometric structural information about the face. The discriminator part consists of a global discriminator and three local discriminators, which serve as a facial attention mechanism. According to facial characteristics, we crop the input face images to three different regions (eyes, nose, and mouth), which have the most discriminative features. Additionally, we add the self-attention block to the uppermost and second-uppermost layers in each discriminator. This, can help capture long-distance dependencies between image features. Moreover, multiple loss functions are exploited in the training process to help synthesize frontal face images that are more similar to the real images.

The main contributions of this paper can be summarized as follows.

• We propose a face frontalization model based on GAN (DMA-GAN), which can synthesize identity-preserved frontal-view face images from multi-pose face images and does not require the input of other prior knowledge of the face, such as the type of pose being captured.

• We combine the residual block with the attention mechanism to form a deep feature encoder, which can deepen the network layers and extract more abstract facial details. We also add the attention mechanism in the discriminator.

• Compared with some existing advanced methods, our model is simpler and achieves higher recognition rates for some angles. The results from quantitative and qualitative experiments prove the effectiveness of the proposed method.

II. RELATED WORK

A. Generative Adversarial Networks

Goodfellow et al. [9] first proposed the generative adversarial networks (GANs). The min-max two-player game provides a simple yet powerful way to estimate target distribution and generate novel image samples [23]. Due to its excellent performance, it has drawn substantial attention and has been widely used in deep learning and computer vision. In order to improve some problems with GAN, such as training instability, researchers have made

various modifications to GANs from the perspective of architecture or loss function. DCGAN [24] applies deep convolution to GAN and achieves significant improvement. WGAN [25] and WGAN-GP [26] use the Wasserstein distance instead of KL-divergence in GAN, which improves the stability of GAN training. BEGAN proposes a new equilibrium enforcing method paired with a loss derived from the Wasserstein distance for training auto-encoder based generative adversarial networks [27]. These methods greatly advance in various generation tasks.

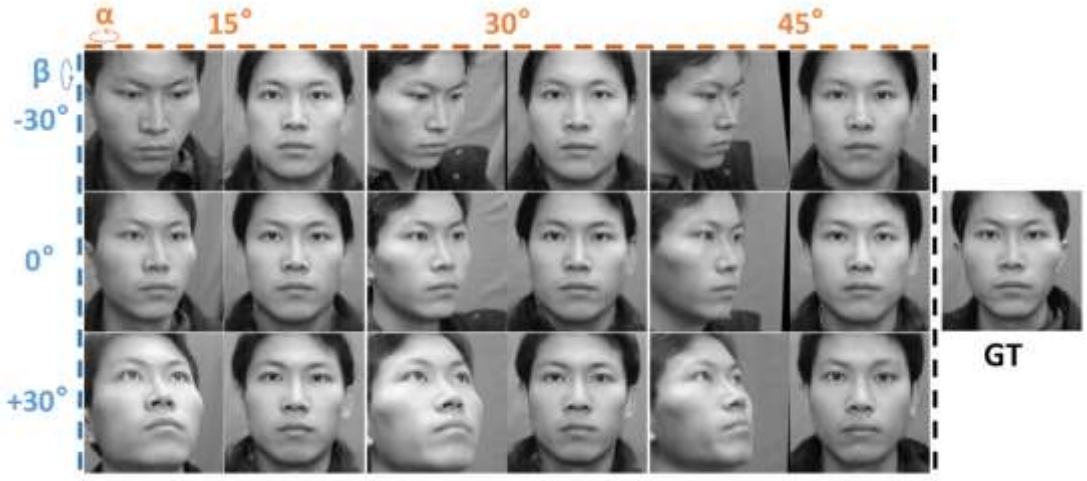

Fig. 1: The synthesized face frontalization results of DMAGAN. The first column of each face pair displays real multi-pose faces from CAS-PEAL-R1, and the second column displays synthesized frontal faces. The ground truth frontal face is shown on the right side.

B. Face Frontalization

Face frontalization is a computer vision task that synthesizes identity-preserved frontal-view faces from various viewpoints. Existing methods addressing the face frontalization problem can be divided into two categories: 2D-based methods [6] [7] [8] and 3D-based methods [28] [29]. Hassner et al. [28] use an unmodified 3D surface as an approximation for all input surface shapes to produce frontal faces images. Zhu et al. [29] propose a pose-adaptive 3DMM-fitting algorithm. The synthetic 3D-based images are not often realistic, and performance deteriorates for large poses. In recent years, the development of GAN has greatly improved the visual effect of the two-dimensional image generation task. Many face frontalization methods based on GAN have been proposed. For instance, Huang et al. [10] propose a deep architecture with two pathways (TP-GAN) focusing on global structure and local texture respectively, for frontal view synthesis. Qian et al. [17] propose a face normalization model (FNM) to synthesize frontal, neutral expressions and photorealistic face images in the condition of an unconstrained environment. Tran et al. [11] propose DR-GAN to solve the face frontalization problem faces by extending GAN with an encoder-decoder structured generator and pose code.

In view of the effectiveness of GAN, our model is also based on GAN.

C. Attention Mechanism

Recently, many researchers have improved the expression of features by adding attention mechanisms to the network. The attention mechanism serves as a way of making the network emphasize regions of interest and suppress regions of irrelevant background through

self-learning, which is essentially similar to the way humans observe things. In 2014, Mnih et al. [30] first used attention in recurrent neural nets for image classification. Then the attention mechanism became widely used in various natural language processing tasks [31] [32]. In recent years, attention mechanisms have also played an important role in computer vision. In particular, the DA-Net [21] aggregates the position attention module and the channel attention module to capture long-range relationships for accurate segmentation. Han et al. [22] introduce the self-attention to GAN (SAGAN), which could produce a more detailed and vivid visualization. For face frontalization tasks, DA-GAN [13] and GSP-GAN [14] place the self-attention module in the generator and discriminator respectively. In our model, we combine and stack the position attention module and channel attention module to produce more abstract features.

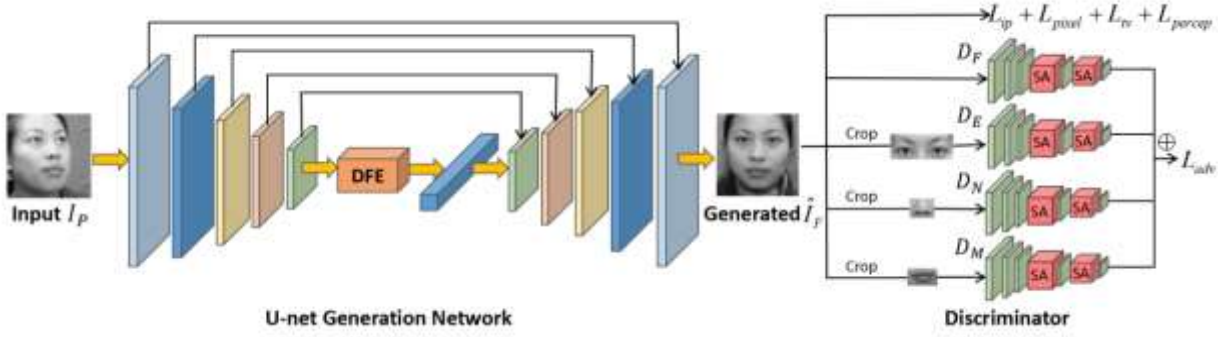

Fig. 2: Framework of DMA-GAN. The generation network is based on U-net architecture and is added with a deep feature encoder (DFE). The discriminator part consists of four independent discriminators.

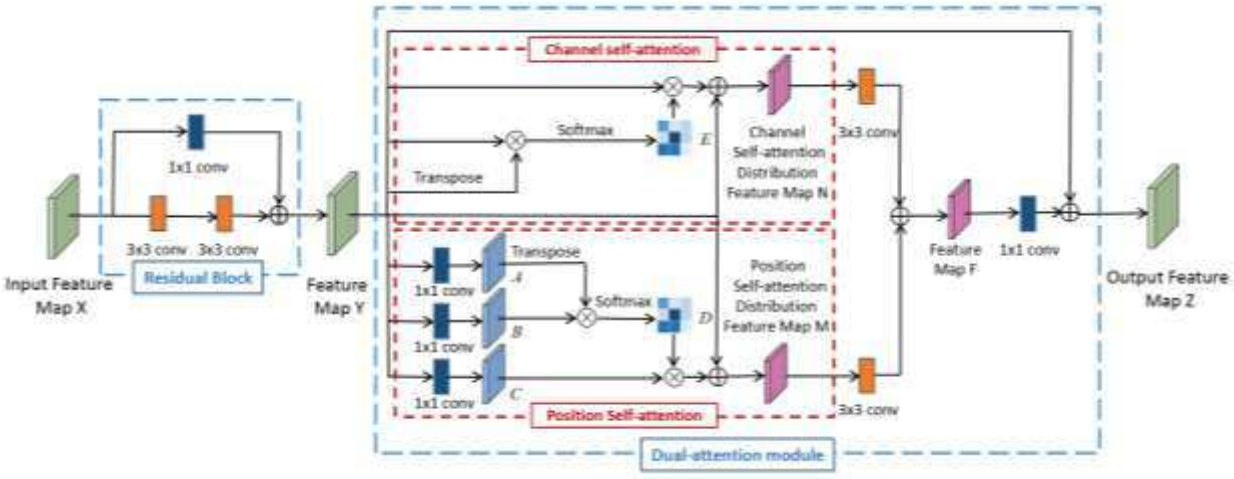

Figure3: The processing flow of the deep feature encoder module. The DFE in the network consists of four stacked modules as shown in this figure. "⊕" represents the matrix element-wise sum, and "⊗" represents matrix multiplication.

III. METHOD

The structure of our model is shown in Fig. 2. The input multi-pose image is denoted as I^P, the corresponding frontal image is denoted as I^F, the synthesized frontal image is denoted as $\hat{I}^F$. The encoder, deep feature encoder and decoder are denoted as G_E, G_{DFE},

G_D. The frontal face is cropped into three regions: eyes I_E; nose I_N; mouth I_M. The corresponding three synthesized regions are denoted as $\hat{I}_E$, $\hat{I}_N$ and $\hat{I}_M$. The discriminators are denoted as D_F, D_E, D_N and D_M.

A. Generator

The generator of our model, shown in Fig. 2 is based on U-net architecture that consists of an encoder-decoder structure for image synthesizing. Skip connections are used between the encoder and decoder to enable multi-scale feature fusion. Inspired by [33], we add a deep feature encoder (DFE) behind the encoder. The generation process can be described as:

$$\hat{I}^F = G_D\left(G_{DFE}\left(G_E\left(I^P\right)\right)\right) \tag{1}$$

The DFE consists of four stacked modules, as shown in Fig. 3. The deep feature encoder module comprises composed two parts: a residual block and dual-attention module. He et al. [34] propose ResNet, which is easier to optimize and can gain accuracy from increased depth. In order to extract more abstract facial features, we stack up basic residual blocks to deepen the generation network. Inspired by [21], we combine position self-attention and channel self-attention into a dual-attention module to capture long-range relationships and model interdependencies between channels. Detailed architecture is shown in Table I.

The input feature map *X* firstly goes through a basic residual block. The position and channel self-attention modules are parallel. As follows, we describe the flow of the two modules respectively.

(1) Position Self-attention Module

Given a feature map $Y \in R^{C\times H\times W}$, we first generate three new feature maps *A*, *B* and *C* by feeding *Y* into three different 1×1 convolutional layers, where $\{A,B,C\} \in R^{C\times H\times W}$. Then, we reshape *A* and *B* to $R^{C\times N}$, where $N = H\times W$. Then we perform matrix multiplication between A^T and *B* and apply a softmax layer to obtain the spatial feature map $D \in R^{N\times N}$:

$$d_{ji} = \frac{\exp\left(B_i \cdot A_j\right)}{\sum_{i=1}^{N}\exp\left(B_i \cdot A_j\right)} \tag{2}$$

where d_{ji} indicates the degree of position *i*'s impact on position *j*. The two positions have more similar feature, which indicates they are more highly correlated.

Meanwhile, we reshape *C* to $R^{C\times N}$. Then, we perform matrix multiplication between C^T and *D* and reshape the result to $R^{C\times H\times W}$. Finally, we multiply the result by a scale parameter α and perform an element-wise sum operation with the original feature *Y,* obtaining the final-position self-attention distribution feature map $M \in R^{C\times H\times W}$:

$$M_j = \alpha \sum_{i=1}^{N} \left(d_{ji} C_i \right) + Y_j \tag{3}$$

where the value of α is initialized to 0 and adapted during training. From Equation 3, we can infer that the feature M of each position is the weighted sum of the features of all positions and the original feature. Therefore, it provides a global contextual view that can help aggregate the feature information of contexts selectively.

(2) Channel Self-attention Module

Given a feature map $Y \in R^{C \times H \times W}$. Unlike with calculating the position attention map, we can directly obtain the channel attention map $E \in R^{C \times C}$ from the original feature map Y without convolutional layers. To do so, we first reshape Y to $R^{C \times N}$ and perform matrix multiplication between Y and Y^T. After that, we apply a softmax layer to obtain the channel attention map $E \in R^{C \times C}$:

$$e_{ji} = \frac{\exp\left(Y_i \cdot Y_j \right)}{\sum_{i=1}^{C} \exp\left(Y_i \cdot Y_j \right)} \tag{4}$$

where e_{ji} indicates the degree of the i^{th} channel's effect on the j^{th} channel. Additionally, we perform matrix multiplication between E and Y and reshape the result to $R^{C \times H \times W}$. Finally, we multiply the result by a scale parameter β and perform an element-wise sum operation with the original feature Y, obtaining the final channel self-attention distribution feature map $N \in R^{C \times H \times W}$:

$$N_j = \beta \sum_{i=1}^{N} \left(e_{ji} Y_i \right) + Y_j \tag{5}$$

where β starts at 0 and learns weight gradually during training. It can be inferred from Equation 5 that the final feature of each channel is the weighted sum of all channel features and the original feature. Therefore, it can enhance the model's ability to distinguish features.

After obtaining position self-attention feature map M and channel self-attention feature map N, we feed them into a 3×3 convolutional layer and perform an element-wise sum operation to obtain feature map F. Finally, we apply a 1×1 convolution operation for F and add the result with the original feature Y to obtain the final feature map Z.

B. Discriminator

Yin et al. [13] parse the frontal face image into three predefined regions (skin, keypoints, and hairline) and assign each region to a regional discriminator. GSP-GAN [14], FR-DVF [16], and FNM [17] use segment strategy in the discriminator by cropping different facial patches. In our model's discriminator, we also use segment strategy to implement a facial attention mechanism by cropping the face image into three regions-eyes, nose, and mouth-which are the most discriminative areas in face recognition. The whole image and these three regions are fed into four independent discriminators (D_F, D_E, D_N, D_M), as

shown in Fig. 2. This strategy can increase the punishment on the generator in order to gain a more realistic frontal face image.

Table 1. Architectures for deep feature encoder (behind the original encoder). The second column presents the structure of the residual block. The third column shows the number of input and output channels of the attention module.

Layer name	Residual block	Attention module channels
Conv5_x	$\begin{bmatrix} 3\times3 & 512,512 \\ 3\times3 & 512,512 \end{bmatrix}$	512
Conv6_x	$\begin{bmatrix} 3\times3 & 512,1024 \\ 3\times3 & 1024,1024 \end{bmatrix}$	1024
Conv7_x	$\begin{bmatrix} 3\times3 & 1024,1024 \\ 3\times3 & 1024,1024 \end{bmatrix}$	1024
Conv8_x	$\begin{bmatrix} 3\times3 & 1024,2048 \\ 3\times3 & 2048,2048 \end{bmatrix}$	2048

Inspired by the use of self-attention in SA-GAN [21], we also add the self-attention block to the uppermost and second-uppermost layers in each discriminator. The self-attention block is the position self-attention block shown in Fig. 3. In addition, we add the residual block after the fourth convolutional layer to deepen the network, so as to extract more abstract facial detail features and improve the discriminant ability.

C. Loss Function

The loss function is a weighted sum of five individual loss functions.

(1) Global-Local Adversarial Loss

The loss for distinguishing real images from synthesized images is the sum of a global loss and three local losses. Mathematically speaking, it is phrased as follows:

$$L_{adv} = \sum_{j\in\{F,E,N,M\}} \left(E_{I_j}\left[\log D_j\left(I_j\right)\right] + E_{\hat{I}_j}\left[\log\left(1 - D_j\left(\hat{I}_j\right)\right)\right]\right) \tag{6}$$

where subscript j represents the facial region and the corresponding label of the discriminator.

(2) Identity Preserving Loss

Preserving the identity is a critical part of synthesizing the frontal face image. We exploit the pre-trained 29-layer LightCNN [35] to give our model the ability to preserve identity. LightCNN is trained on a large-scale face dataset to enable it to extract more general and more significant facial features. The identity-preserving loss is denoted as:

$$L_{ip} = \sum_{i=1}^{2} \left\| \varphi_i\left(I^F\right) - \varphi_i\left(\hat{I}^F\right) \right\|_2^2 \tag{7}$$

where $\varphi(\cdot)$ is the output feature from the fully connected layers of pre-trained Light CNN and $\|\cdot\|_2$ is the L2-norm.

(3) Multi-scale Pixel-wise Loss

Following [12], we employ a multi-scale pixel-wise loss to constrain the content consistency between the synthesized $\hat{I}^F$ and the corresponding frontal image I^F. Mathematically speaking:

$$L_{pixel} = \frac{1}{3}\sum_{i=1}^{3}\frac{1}{W_i H_i C}\sum_{w,h,c}^{W_i,H_i,C}\left|\hat{I}^F_{i,w,h,c} - I^F_{i,w,h,c}\right| \tag{8}$$

where C is the channel number, and W_i, H_i are the corresponding width and height of the i^{th} scale. The scales are image size: 128×128, 64×64, 32×32.

(4) Perceptual Loss

Because pixel-level errors can not capture the perceived difference between the real image and the synthesized image, Johnson et al. [36] propose using perceptual loss to measure the similarity between images. We use pre-trained vgg19 [37] as the feature extractor to gain feature maps of the real image and the synthesized image. This technique can help the image to be generated in a way that is more semantically similar to the target image by comparing the feature maps. Mathematically speaking, this can be expressed as follows:

$$L_{percep} = \frac{1}{HWC}\left\|\varphi\left(\hat{I}^F\right) - \varphi\left(I^F\right)\right\|_2^2 \tag{9}$$

where C is the channel number, and W, H are the width and height of the feature map.

(5) Total Variation Regularization

We introduce the total variation regularization [36] to remove artifacts and improve the synthesis quality of images, this is written as:

$$L_{tv} = \sum_{c=1}^{C}\sum_{w,h=1}^{W,H}\left|\hat{I}^F_{w+1,h,c} - \hat{I}^F_{w,h,c}\right| + \left|\hat{I}^F_{w,h+1,c} - \hat{I}^F_{w,h,c}\right| \tag{10}$$

where C, W and H are the channel number, width and height of the synthesized image $\hat{I}^F$.

(6) Overall Objective Function

The final objective loss function is a weighted sum of all the aforementioned losses:

$$L_{syn} = \lambda_1 L_{adv} + \lambda_2 L_{ip} + \lambda_3 L_{pixel} + \lambda_4 L_{percep} + \lambda_5 L_{tv} \tag{11}$$

where λ_1, λ_2, λ_3, λ_4 and λ_5 are hyper-parameters corresponding to each loss term.

IV. EXPERIMENT

A. Experiment Settings

(1) Datasets

The CAS-PEAL-R1 dataset [38] is a large public released Chinese face database. It contains 1040 subjects, including 595 males and 445 females, and it includes more than 30000 grayscale images under pose, expression and lighting variations. We only use images with various poses to verify the performance of our face frontalizaiton model. Variations of pose encompass 21 yaw-pitch rotations, including 6 yaw angles (i.e., α = {0◦, ±15◦, ±30◦, ±45◦}), 3 pitch angles (i.e., β = {0◦, ±30◦}). In our work, the first 600 subjects are used for training and the remaining 440 subjects are used for testing.

CASIA-FaceV5 [39] is a large colorful Asian face database collected by the Chinese Academy of Sciences' Institute of Automation. It contains 2,500 color facial images of 500 subjects. The intra-class variations include illumination, pose, expression, eye-glasses, imaging distance differences. We use images of poses for the testing experiment.

(2) Implementation Details

In the training process, we use pairs of frontal and non-frontal face images $\{I^F, I^P\}$ from CAS-PEAL-R1 dataset for input. We reshape all images to a canonical view of size 128×128, and both real and generated images are grayscale images. The identity-preserving network is pre-trained on grayscale images from MS-Celeb-1M. In the training process, we use the Adam optimizer to train the generator and the discriminators, and the parameters of the network are updated alternately. In our experiments, we set the hyperparameters of the objective function as: $\lambda_1 = 1.0$, $\lambda_2 = 0.01$, $\lambda_3 = 10$, $\lambda_4 = 0.01$, $\lambda_5 = 0.01$. We implemented our network with PyTorch.

(3) Qualitative Results

To verify the performance of our model, we conduct test experiments on two datasets respectively. For the CAS-PEAL-R1 dataset, Fig. 4 shows the DMA-GAN's ability to synthesize frontal identity-preserving face images from various perspectives. In multi-pose face recognition tasks, a wide-angle pose presents a very challenging problem. DMA-GAN can synthesize high-quality frontal face images that maintain clarity of facial features and stable facial structure, even when much semantic information is lost due to wide-angle perspectives (β = -30°, α = 45°).

To further demonstrate the high performance of our model, we also visually compare synthesized images produced by CAS-PEAL-R1 with state-of-the-art methods (TP-GAN [10], CR-GAN [18], M^2FPA [12], DA-GAN [13]), as shown in Fig. 5. These synthesis images qualitatively demonstrate the effectiveness of our model (DMA-GAN). These synthesized images qualitatively demonstrate the effectiveness of our model (DMA-GAN). We can observe that DMA-GAN displays strong performance in both facial texture detail and geometric shape.

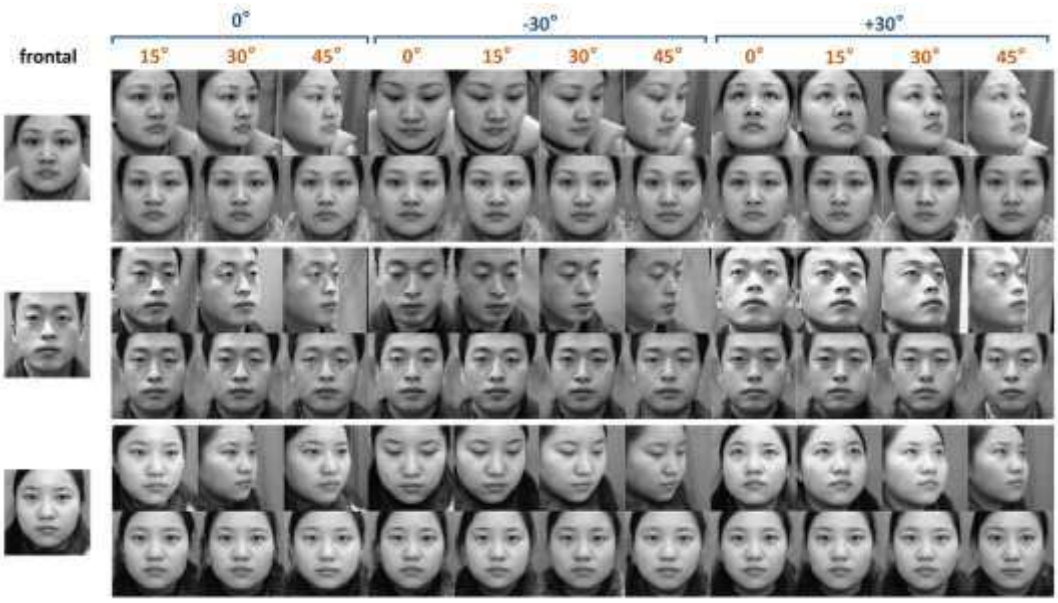

Figure 4: Synthesis results by DMA-GAN for various poses. The first line of labels are pitch angles (i.e., β = {0◦, ±30◦}), and the second line of labels are yaw angles (i.e., α = {0◦, 15◦, 30◦, 45◦}). We display a group of images for the same person in 11 poses and the synthesized frontal images are presented below.

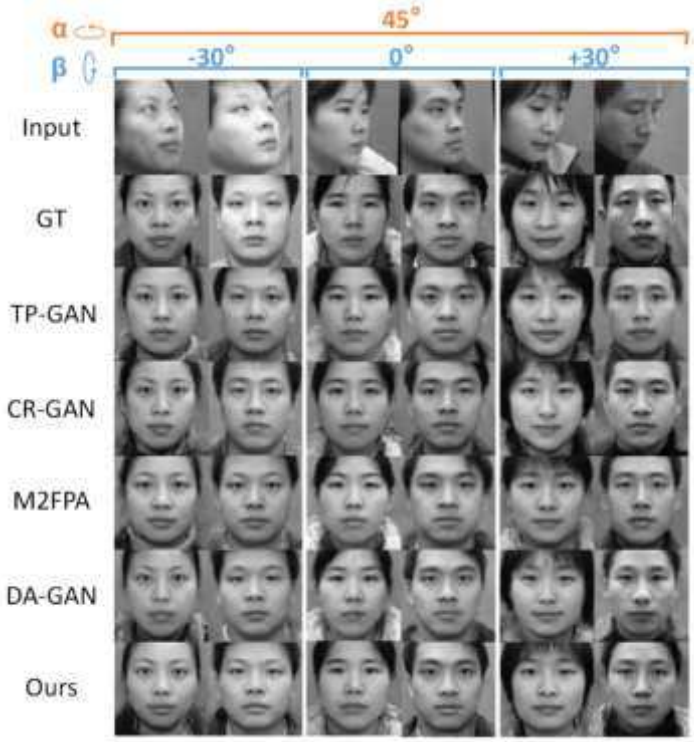

Figure 5: Synthesis results. Comparison with several popular methods on yaw (α) and varying pitch (β) angles.

To demonstrate our model's generalization ability, we use images from CASIA-FaceV5 dataset to test our model trained solely on CAS-PEAL-R1. Because our model is trained with grayscale images, we first convert images from CASIA-FaceV5 to grayscale. The synthesis results, shown in Fig. 6, reveal that our model can faithfully synthesize frontal-view face images.

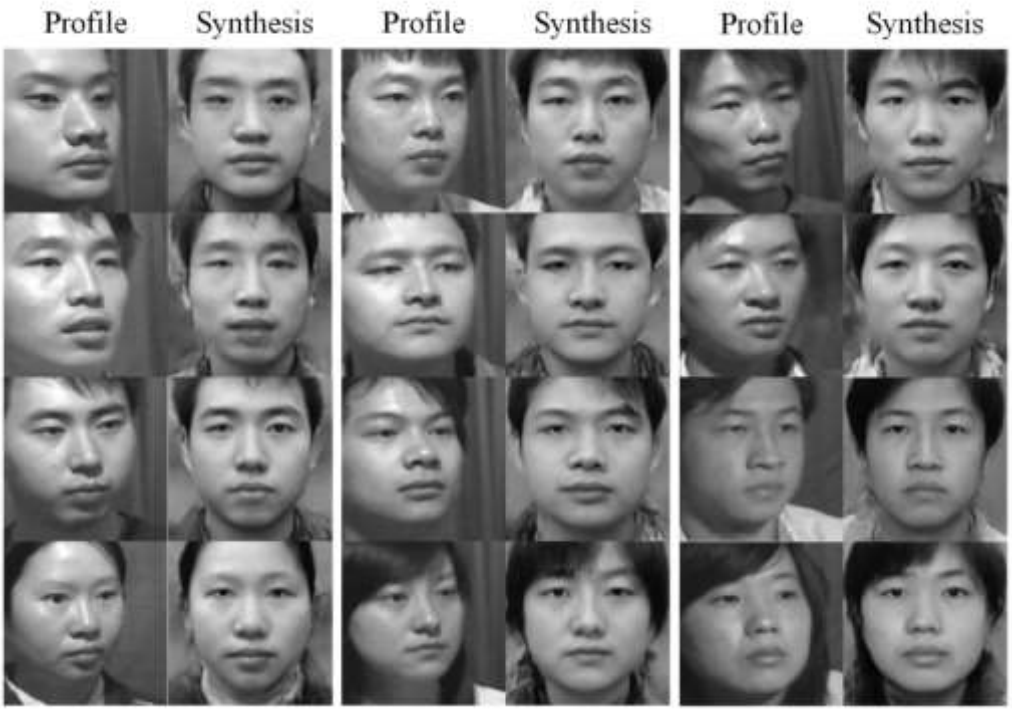

Figure 6: Synthesis results on CASIA-FaceV5 dataset. DMA-GAN is trained on CAS-PEAL-R1.

(4) Quantitative Results

We conduct face recognition with the CAS-PEAL-R1 dataset to quantitatively verify the identity-preserving ability of DMA-GAN. We follow the open evaluating protocol, and use a pre-trained 29-layer Light-CNN [35] to extract deep features and cosine-distance metric to calculate similarity. We test the rank-1 recognition rate on the CAS-PEAL-R1 dataset. Table II shows the recognition rate of our model and some popular methods.

Table II : Rank-1 recognition rates (%) across yaw (α) and pitch (β) pose variations under CAS-PEAL-R1.

	Pitch(-30°)					Pitch(0°)				Pitch(+30°)				
Yaw	±0°	±15°	±30°	±45°	Avg_1	±15°	±30°	±45°	Avg_2	±0°	±15°	±30°	±45°	Avg_3
TP-GAN	98.86	98.94	98.89	97.62	98.58	100.00	99.94	98.71	99.55	97.68	97.73	97.45	95.83	97.17
CR-GAN	83.98	83.91	83.17	80.38	82.86	97.61	95.80	89.73	94.38	89.74	89.44	87.95	83.90	87.76
M2FPA	99.38	99.42	99.30	98.53	99.16	100.00	99.94	99.36	99.77	98.60	98.69	98.58	97.84	98.43
DA-GAN	99.71	99.72	99.65	98.99	99.52	100.00	100.00	99.70	99.90	98.96	98.88	98.86	98.13	98.73
DMA-GAN	96.55	95.59	94.91	91.53	94.65	**100.00**	99.89	97.16	99.02	**99.78**	**98.98**	96.59	93.18	97.13

It can be observed that DMA-GAN outperforms CR-GAN and has the highest recognition rate at some angles (in bold within Table II). In contrast to TP-GAN, M^2FPA and DA-GAN in network structure, TP-GAN uses a two-pathway generative adversarial network which greatly reduces the training efficiency. M^2FPA and DA-GAN use a pre-trained model as a face parse to generate masks which are used as inputs to complete the attention mechanism of the discriminator. However, our model uses a single generation path and does not require

additional input except pairs of frontal and non-frontal face images $\{I^F, I^P\}$. Thus, we achieve not the best but also very efficient performance through a simpler network.

(1) Ablation Study

To verify the superiority of DMA-GAN as well as the contributions of various components, we remove each component individually and test the synthesis performance. The process of training is the same. Our baseline model only has two single parts: a generator based on U-net and an ordinary face discriminator. So we train four partial variants of DMA-GAN: one without deep feature encoder (DFE); one without three discriminators that are used as facial attention mechanism (sub-D); one without self-attention in discriminators (D-att); and one without identity preserving loss (id). The synthesis results of the ablative comparison are shown in Fig. 7. Table III shows the quantitative comparison.

First, the synthesis images without L_{ip} can not maintain the identity-preserved feature well. Second, the model without sub-D and D-att can hardly capture the details of facial features, especially for the eyes, nose and mouth regions, which are the most discriminative areas in face recognition. Moreover, the comparison between the third column and the sixth column demonstrates that the deep feature encoder (DFE) we designed displays notable performance in perceiving local texture and synthesizing more vivid facial details.

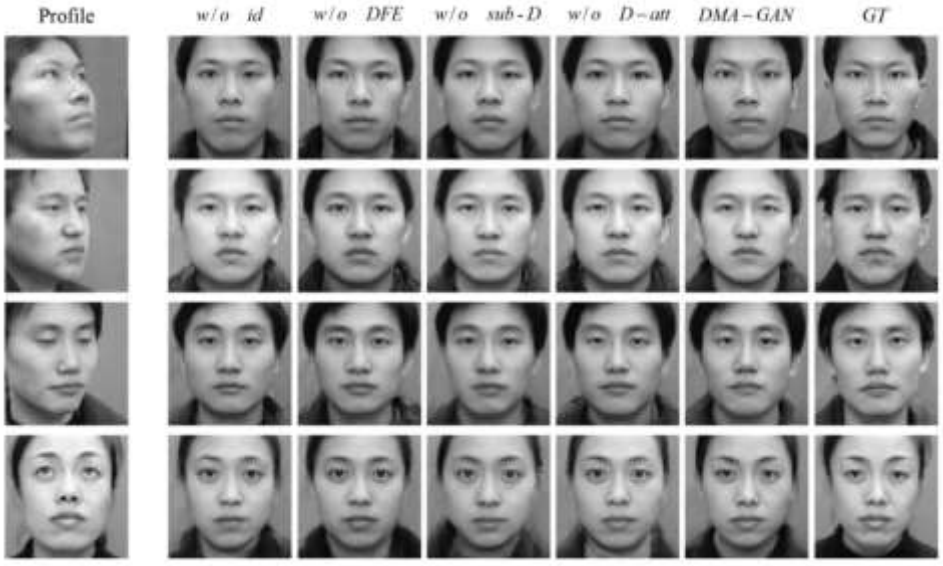

Figure 7: Synthesis results of ablative comparison for DMA-GAN.

Table III: Ablation study: quantitative results. Rank-1 recognition rates (%).

	Pitch(-30°)					Pitch(0°)				Pitch(+30°)				
Yaw	±0°	±15°	±30°	±45°	Avg_1	±15°	±30°	±45°	Avg_2	±0°	±15°	±30°	±45°	Avg_3
w/o DFE	88.81	86.61	83.22	75.25	83.45	91.78	89.75	81.78	87.77	86.69	85.17	81.78	75.17	82.20
w/o sub-D	76.95	75.59	72.20	70.34	73.77	81.69	79.66	77.28	79.54	75.25	76.61	76.44	72.37	75.17
w/o D-att	86.95	84.24	83.56	75.59	82.59	95.42	93.90	86.10	91.80	90.85	86.44	80.85	73.56	82.93
DMA-GAN	**96.55**	**95.59**	**94.91**	**91.53**	**94.65**	**100.00**	**99.89**	**97.16**	**99.02**	**99.78**	**98.98**	**96.59**	**93.18**	**97.13**

V. CONCLUSION

In this paper, we propose a deep GAN-based model which combining multi-attention mechanisms (DMA-GAN) can be effectively used for face frontalization. The proposed model can synthesize high-quality frontal-view face images from multi-pose face images and does not require inputs of other prior knowledge of the face, such as the type of pose represented in the image. For the generator based on U-net, we introduce a depth feature encoder composed of residual-blocks dual-attention modules to capture more abstract detailed features. The discriminator combines the facial attention mechanism and self-attention module. Quantitative and qualitative results demonstrated the validity of our model. The synthesized results are compelling, showing that our model has practical significance.

Recently, many methods have attempted to address the problem of face frontalization in the wild. For instance, CCFF-GAN [15] uses semi-supervised learning to improve the capacity for generalization ability in the unconstrained environment. This is a challenging area that we need to investigate further.

Declarations

Ethical Approval

not applicable

Competing interests

not applicable

Authors' contributions

Jiaqian Cao, Zhenxue Chen, Yujiao Zhang and Luna Sun wrote the main manuscript text, and Jiyang Chen prepared figures 2-3. All authors reviewed the manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China (61876099) in part by the National Key R&D Program of China (2019YFB1311001), and in part by the Key R&D Project of Shandong Province (2022CXGC010503).

Availability of data and materials

"CAS-PEAL-R1" http://www.jdl.ac.cn/peal/

"Casia-facev5" http://biometrics.idealtest.org/

REFERENCES

[1] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in British Machine Vision Conference, 2015.

[2] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1891-1898.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.
[4] K. Cao, Y. Rong, C. Li, X. Tang, and C. C. Loy, “Pose-robust face recognition via deep residual equivariant mapping,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5187–5196.
[5] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, “Pose-guided photorealistic face rotation,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8398–8406.
[6] J. Yim, H. Jung, B. I. Yoo, C. Choi, and J. Kim, “Rotating your face using multi-task deep neural network,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[7] Z. Zhu, P. Luo, X. Wang, and X. Tang, “Multi-view perceptron: A deep model for learning face identity and view representations,” in Advances in Neural Information Processing Systems, 2014.
[8] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, “Synthesizing normalized faces from facial identity features,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3386–3395.
[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Neural Information Processing Systems, 2014.
[10] 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 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2458–2467.
[11] L. Tran, X. Yin, and X. Liu, “Disentangled representation learning gan for pose-invariant face recognition,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1283–1292.
[12] 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 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10 042–10 050.
[13] Y. Yin, S. Jiang, J. P. Robinson, and Y. Fu, “Dual-attention gan for large-pose face frontalization,” in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 249–256.
[14] 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. 104 676–104 687, 2020.
[15] Z. Zhang, R. Liang, X. Chen, X. Xu, G. Hu, W. Zuo, and E. R. Hancock, “Semi-supervised face frontalization in the wild,” IEEE Transactions on Information Forensics and Security, vol. 16, pp. 909–922, 2021.
[16] C. S. D. Q. Luo, H. and C. X., “Frontal face reconstruction based on detail identification, variable scale self-attention and flexible skip connection.” in Neural Comput Applic, 2022.
[17] Y. Qian, W. Deng, and J. Hu, “Unsupervised face normalization with extreme pose and expression in the wild,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9843–9850.
[18] 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 2207–2216.
[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in Computer Vision-ECCV 2018, 2018, pp. 3–19.

[20] Y. A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim, “Unsupervised attention-guided image to image translation,” in Proc. of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 3697–3707.
[21] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3141–3149.
[22] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” 2018.
[23] E. Denton, S. Chintala, A. Szlam, and R. Fergus, “Deep generative image models using a laplacian pyramid of adversarial networks,” in International Conference on Neural Information Processing Systems, 2015, p. 1486–1494.
[24] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” Computer ence, 2015.
[25] M. Arjovsky, S. Chintala, Lxe, and O. Bottou, “Wasserstein generative adversarial networks,” in Proceedings of the 34th International Conference on Machine Learning (ICML), 2017, pp. 214–223.
[26] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, “Improved training of wasserstein gans,” pp. 5767–5777, 2017.
[27] D. Berthelot, T. Schumm, and L. Metz, “Began: Boundary equilibrium generative adversarial networks,” arXiv, 2017.
[28] T. Hassner, S. Harel, E. Paz, and R. Enbar, “Effective face frontalization in unconstrained images,” in Computer Vision Pattern Recognition, 2015, pp. 4295–4304.
[29] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, “High-fidelity pose and expression normalization for face recognition in the wild,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 787–796.
[30] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, “Recurrent models of visual attention,” Advances in Neural Information Processing Systems, vol. 3, pp. 2204–2212, 2014.
[31] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” Computer Science, 2014.
[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in arXiv, 2017.
[33] S. Duan, Z. Chen, Q. Wu, L. Cai, and D. Lu, “Multi-scale gradients self-attention residual learning for face photo-sketch transformation,” IEEE Transactions on Information Forensics and Security, vol. 16, pp. 1218–1230, 2021.
[34] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[35] X. Wu, R. He, Z. Sun, and T. Tan, “A light cnn for deep face representation with noisy labels,” IEEE Transactions on Information Forensics and Security, vol. 13, no. 11, pp. 2884–2896, 2018.
[36] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution”,” in Computer Vision – ECCV 2016”, 2016, pp. 694–711.
[37] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” Computer Science, 2014.
[38] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, “The cas-peal large-scale chinese face database and baseline evaluations,” IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 38, no. 1, pp. 149–161, 2008.

[39] “Casia-facev5,” http://biometrics.idealtest.org/

Jiaqian Cao was born in Hebei, China, in 1998. She received the B.S. degree in automation from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2020, where she is currently pursuing the M.S. degree in control science and engineering. Her research interests include machine learning, deep learning, and face frontalization.

Zhenxue Chen was born in Shandong, China, in 1977. He received the B.S. degree in automatic from the School of Electrical Engineering and Automation, Shandong Institute of Light Industry, Jinan, China, in 2000, the M.S. degree in computer science from the School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, China, in 2003, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Image Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, in 2007. From 2012 to 2013, he was a Visiting Scholar with Michigan State University, East Lansing, MI, USA. He is currently a Professor with the School of Control Science and Engineering, Shandong University. He has published over 100 papers in refereed international leading journals/conferences, such as IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, Information Sciences, Neurocomputing, Neural Computing and Applications, and SP-IC. His main areas of interest include image processing, pattern recognition, and computer vision, with applications to face recognition.

Yujiao Zhang was born in Anhui, China, in 1999. She received the B.S. degree in automation from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2020, where she is currently pursuing the M.S. degree in control science and engineering. Her research interests include machine learning, deep learning, and face super-resolution.

Luna Sun was born in Henan, China, in 1996. She received the B.S. degree in automatic from the School of Internet of Things Engineering, Jiangnan University, Wuxi, China, in 2019 and the M.S. degree in Control Engineering from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2022. She will begin to pursue the Ph.D. degree at the School of Control Science and Engineering, Shandong University, Jinan, China, from September 2022. Her current research interests include machine learning, deep learning, and salient object detection.

Jiyang Chen was born in Shandong, China, in 1991. He received the B.S. and M.S. degrees in automation from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2013 and in 2017, where he is currently pursuing the Ph.D. degree in control science and engineering. His research interests include machine learning, deep learning, and face recognition.