IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Frontal Face Generation Based on Attitude Point Guidance and Attention Mechanism Improvement

JIHUI ZHAO[1], ZHENGYI YUAN[2] (Student Member, IEEE), JUNHU ZHANG[3], HAITAO LI[4], HUI LI[5]

[1]School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao Shandong 266000 China (e-mail:1577198220@qq.com)
[2]School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao Shandong 266000 China (e-mail:yuanzhengyi0224@163.com)
[3]School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao Shandong 266000 China;
[4]School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao Shandong 266000 China;
[5]School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao Shandong 266000 China;

Corresponding author: Third C. JUNHU ZHANG (e-mail: jzhang@qust. edu.cn).

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61806107, 61702135 and 62201314, the Opening Project of State Key Laboratory of Digital Publishing Technology, and the Shandong Province "Double-Hundred Talent Plan" on 100 Foreign Experts and 100 Foreign Expert Teams Introduction (WST2021020).

ABSTRACT The deflection of face angle is the most important factor affecting the accuracy in face recognition, non-frontal faces make some face recognition systems lose their due functions,the existing frontal face conversion methods often have the phenomenon of distortion and lack of identity. Aiming at the above problems, this paper proposes an face-frontal network which combines heat map of key points of face and improved attention mechanism. The network consists of a generator network and two discriminators networks, and the thermal map of the key points of the frontal face is used as a priori condition to guide the generation of the frontal face. In the generator part, the self-attention mechanism is introduced to obtain the dependence between feature points and other position features, which enhances the illumination perception ability of the network layer. At the same time, local attention is used in a discriminator to improve the local detail generation ability of the network in the face. Compared with other advanced frontal face generation methods, the proposed method has improved the accuracy of Rank-1 face recognition compared with other methods. The recognition rate of Rank-1 on Multi-PIE data set with small angle deflection is higher than other methods, and the average recognition rate of Rank-2 is 97.41%, which is higher than the advanced PIM method. Experimental results show that the proposed method can generate positive faces with corresponding identities from non-positive faces, which can be directly used in recognition tasks and has high recognition accuracy.

INDEX TERMS Frontal face,Generation model,Attention mechanism,Face identification.

I. INTRODUCTION

THE generalization capability of face recognition models is proportional to the scale of the training data [1]. While the accuracy is high for standard frontal faces, performance declines in uncontrolled scenarios like profile faces, indicating that pose variation is a major factor affecting recognition rates. Current research on this issue can be categorized into two main types. The first type involves directly extracting pose-invariant facial features for recognition tasks [2], [3]. These methods utilize metric learning to achieve pose-invariant feature embeddings; however, the unbalanced nature of pose variations complicates the attainment of ideal pose-invariant features. Furthermore, multi-pose face recognition requires retraining the model each time, adding to the complexity of the process.

The second type of method synthesizes side-profile faces into frontal standard faces for recognition tasks. Zhang et al [4]. extracted local Gabor magnitude binary pattern features from side-profile images to create feature histograms, connecting corresponding mapping histograms to generate frontal faces. 3D model-based methods [5]–[7] adjust facial angles to some extent, but the synthesized faces exhibit noticeable artificial artifacts, and images at high angles may lose authenticity.

In recent years, Generative Adversarial Networks (GANs) [8] have drawn considerable attention due to their powerful image generation capabilities. Methods for generating facial images based on generative models produce more realis-

tic results than those generated by 3D methods, reducing model complexity and computational load. Huang et al. [13] proposed a Two-Pathway Generative Adversarial Network (TP-GAN), which processes global facial contours and local organ details through two distinct pathways. CAPG-GAN [9] (Pose-Guided Photorealistic Face Rotation) employs pose point embeddings and dual discriminators to guide the generation of faces from multiple angles. Xu Haiyue [28] introduced a multi-pose face frontalization approach based on an encoder-decoder architecture. The Unsupervised Normalization Model (Face Normalization Model, FNM) [29] enhances model generalization by utilizing a pre-trained face feature extraction network and employs multiple local discriminators to increase penalties on the generator, generating frontal faces with a uniform style. Xin Jingwei et al [30]. designed a hierarchical representation integration inference network that synthesizes frontal faces by combining low-level visual information with high-level semantic information without introducing prior knowledge.

However, the aforementioned methods based on encoder-decoder structures or GANs struggle to capture global features, as the nature of convolution can only capture local information and fails to address long-range feature dependencies. Additionally, a singular discriminator cannot effectively supervise the generator's output in greater detail. This paper proposes a face frontalization model that combines pose point guidance with a multi-attention mechanism, based on deep learning and Generative Adversarial Networks (GANs). A standard frontal facial coordinate map is selected as the pose prior condition. The self-attention mechanism enhances the generator's ability to process global features, while the local attention mechanism improves the discriminator's supervisory capability. The main contributions include: (1) Utilizing a standard single-channel frontal facial landmark map as prior pose information, combined with a multi-attention mechanism to guide the generation of frontal faces. (2) Incorporating a self-attention module in the decoder to capture long-range pixel dependencies, enhancing the generated images' adaptability to lighting conditions and overall quality. The discriminator employs a local attention mechanism to supervise each local facial area separately, further improving facial detail. (3) Training and testing results on the CAS-PEAL-R1 and Multi-PIE datasets indicate that this method can generate realistic frontalized faces for preprocessing in profile face recognition, enhancing recognition capability. Qualitative and quantitative experiments demonstrate that the proposed method achieves excellent performance in facial frontalization.

II. RELATED WORK

A. GENERATIVE ADVERSARIAL NETWORKS AND FACE FRONTALIZATION

The GAN model proposed by Ian Goodfellow employs a generator and a discriminator in adversarial learning, continuously enhancing the learning capabilities of both until the training concludes when the discriminator can no longer distinguish between the generator's outputs and the real inputs. Mirza and Osindero [10] introduced conditional variables c into GANs to control the direction of the generator's outputs. Arjovsky et al [11]. proposed WGAN to replace the original Jensen-Shannon divergence, allowing the calculation of loss even when the two image distributions do not overlap, thus mitigating the vanishing gradient problem. To stabilize the training process, Gulrajani et al [12]. introduced WGAN-GP, which utilizes a gradient penalty term to constrain the discriminator's parameters, ensuring that the model satisfies the Lipschitz condition.

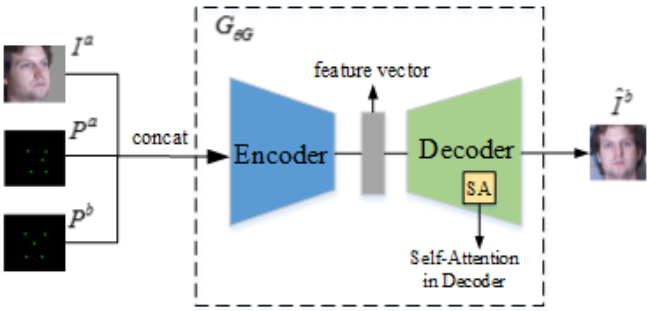

FIGURE 1. Generator of Model

GANs have also made significant advancements in generating frontal faces. The Two-Pathway GAN (TP-GAN) [13] utilizes a dual-pathway architecture that combines global facial contours and local information to generate frontal faces. Zhao et al [14]. expanded upon TP-GAN by introducing a domain adaptation strategy in the PIM model to address facial recognition tasks under extreme poses. FF-GAN [15] integrates 3D techniques with the GAN model, training on 3D models within the GAN framework to synthesize frontal faces from non-frontal faces captured at extreme angles.

B. POSE POINT GUIDANCE AND ATTENTION

The TP-GAN and CAPG-GAN models utilize facial landmark maps to guide the generation of multi-pose faces. The objective of this study is to generate frontal face views, so the average positions of landmarks for frontal faces are taken as the target pose guidance map. The MTCNN [24] is employed to detect the positions of five key facial landmarks from the input face, which are then output as a keypoint heatmap.

The attention mechanism attempts to learn from human vision during the image perception process, where human perception typically focuses on regions of interest. Attention was first utilized in image classification with recurrent neural networks [16] . Zhang et al [17]. were the first to incorporate a self-attention mechanism into the GAN model, enhancing the model's ability to process global information and improve image generation quality. The self-attention mechanism captures dependencies at different positions of a single sequence, compensating for the limitations in feature correlation caused by convolution operations. It has been shown to have a positive impact in computer vision. In this study, a self-attention

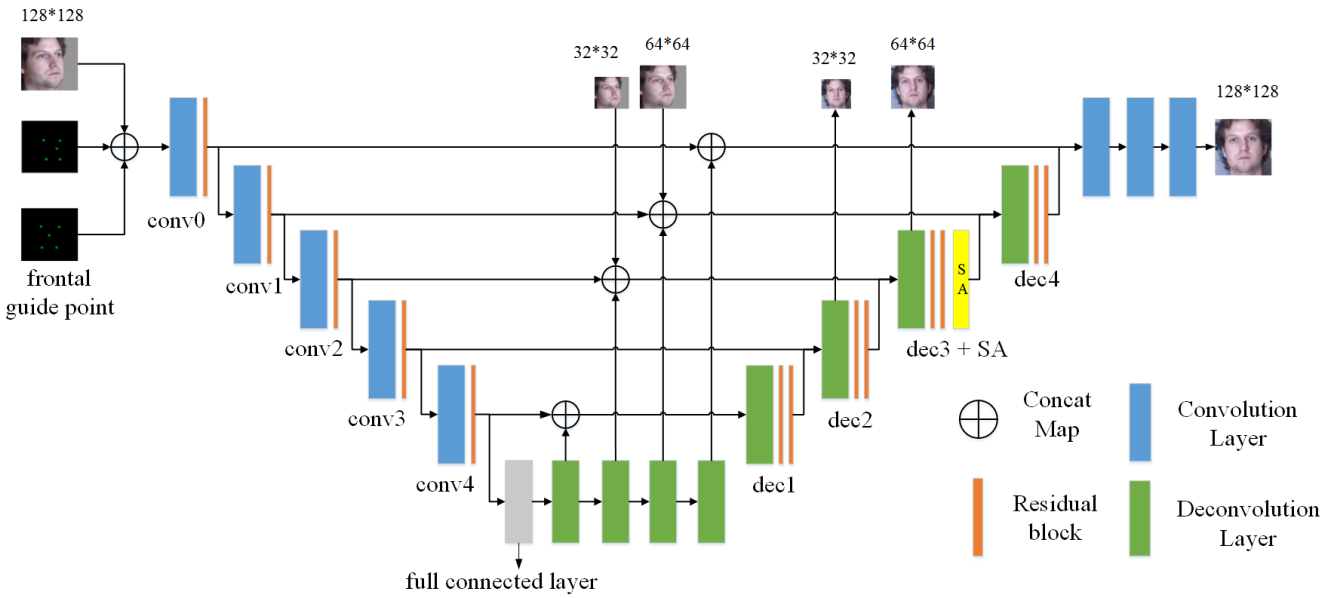

FIGURE 2. Encoder and decoder of generator

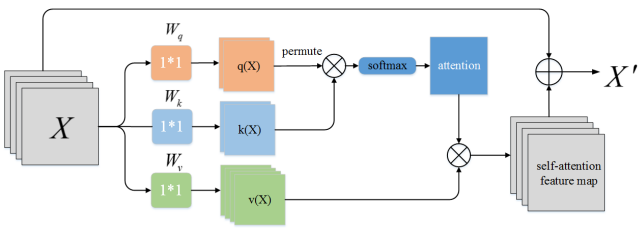

FIGURE 3. Self-attention feature map process

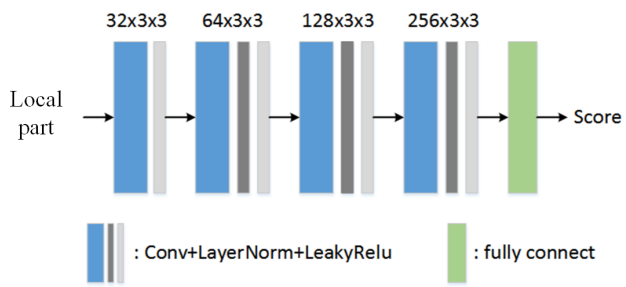

FIGURE 4. Local discriminator structure

module is applied to optimize the generator, while a local attention mechanism enhances the discriminator.

III. METHOD

The TP-GAN and CAPG-GAN models utilize facial landmark maps to guide the generation of multi-pose faces. The objective of this study is to generate frontal face views, so the average positions of landmarks for frontal faces are taken as the target pose guidance map. The MTCNN [24] is employed to detect the positions of five key facial landmarks from the input face, which are then output as a keypoint heatmap.

The attention mechanism attempts to learn from human vision during the image perception process, where human perception typically focuses on regions of interest. Attention was first utilized in image classification with recurrent neural networks [16]. Zhang et al [17]. were the first to incorporate a self-attention mechanism into the GAN model, enhancing the model's ability to process global information and improve image generation quality. The self-attention mechanism captures dependencies at different positions of a single sequence, compensating for the limitations in feature correlation caused by convolution operations. It has been shown to have a positive impact in computer vision. In this study, a self-attention module is applied to optimize the generator, while a local attention mechanism enhances the discriminator.

A. GENERATOR NETWORK STRUCTURE

In this study, frontal keypoint heatmaps are utilized for synthesizing frontal poses, using a face detector [24] to obtain keypoint locations. Keypoints for profile faces and the prepared standard frontal keypoints are collected, resulting in two single-channel landmark heatmaps. These are then combined with the input three-channel face image at the channel level to form a five-channel feature map, which is fed into the generator for training. As illustrated in Figure 1, the original profile image I^a, the original pose landmark map P^a, and the frontal pose landmark map P^b are stacked along the channel dimension before being input into the encoder to obtain features. These features are then sent to the decoder to generate the frontal image $\hat{I}^b$.

The input and output function expressions of the generator

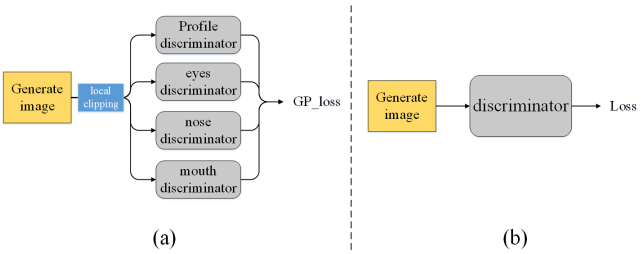

FIGURE 5. (a) The local discriminator (b) Original discriminator

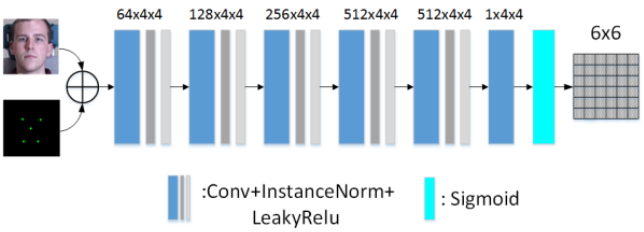

FIGURE 6. D2 discriminator process

are shown in Equations (1) and (2):

$$feature = G_{encoder}\left(I^a + P^a + P^b\right) \tag{1}$$

$$\hat{I}^b = G_{decoder}\left(feature\right) \tag{2}$$

The generator consists of two parts: an encoder and a decoder. The encoder extracts high-dimensional features of the face, primarily for downsampling tasks. The decoder reconstructs the frontal face corresponding to the identity from the low-dimensional feature vectors. Skip connections are used between the encoder and decoder to integrate multi-scale features, effectively utilizing information from different scales. The model components and network structure are shown in Figure 2.

B. THE IMPROVED ENCODER SECTION

Due to the use of local filters in convolutional neural networks, their receptive fields are limited, making it difficult to capture dependencies between distant pixels effectively. To address this limitation, we incorporate self-attention before the penultimate deconvolutional output layer in the decoder section. The self-attention mechanism can directly compute the relationships between any two positions in the feature map, capturing global contextual information and enhancing feature representation.

The core of the self-attention layer is to compute the dependencies between pixels at distant positions in the feature map and apply a nonlinear transformation. By introducing the self-attention mechanism, our model can better capture global features, enhancing the quality of the generated images. The process for computing self-attention feature maps is illustrated in Figure 3. Non-local operations can directly compute

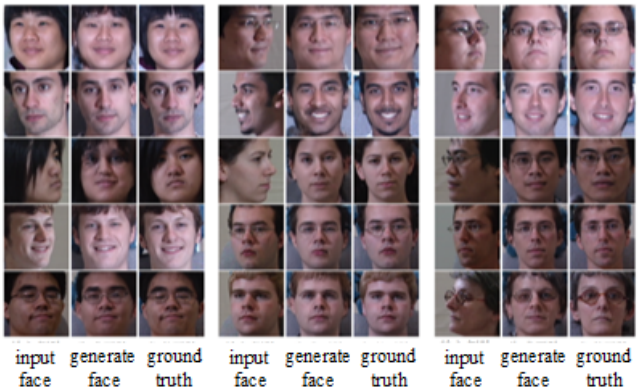

FIGURE 7. Results of different epoch on Multi-PIE training process

FIGURE 8. Frontalization of Faces at Different Angles for the Same Identity

the relationship between two positions in an image, ignoring spatial position influences. The calculation is as follows:

$$y_i = \frac{1}{c(x)} \sum_{\forall j} f\left(x_i, x_j\right) g\left(x_j\right) + x_i \tag{3}$$

Given an input image, the attention map is obtained by multiplying it with the feature map. The weights of the feature map are then normalized using the softmax function. Simultaneously, the original features are input into a convolutional layer, reshaping the dimensions to yield a new feature map. This new feature map is multiplied by the attention weights and then standardized before being added to the original feature map to produce the final result, where represents the position of the feature map. In this paper, a self-attention layer is added after the second-to-last deconvolution layer, dec3. The output of dec3 is a $128\times64\times64$ feature map. After the self-attention layer, the original feature map is added to the self-attention feature map to form the input of dec4, as shown in the decoder section of Figure 2.

C. THE IMPROVED DISCRIMINATOR SECTION

The discriminator section also utilizes two models, D1 and D2. Unlike the use of cross-entropy for loss calculation, this paper employs the gradient penalty strategy of WGAN-GP[9] to optimize the discriminator, stabilizing the entire training process. Considering the local specificity of facial data, we introduce local attention improvements based on the whole

FIGURE 9. Frontalization in smiling expression

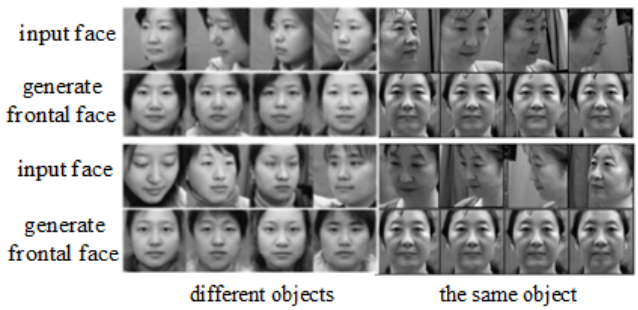

FIGURE 10. Frontalization of CAS-PEAL-R1 data set

face. The input of discriminator D1 is divided into four local regions: face, eyes, nose, and mouth, each fed into corresponding local discriminator sections. The inputs and outputs are depicted in Figure 5.

D1 is composed of four local discriminators, each constructed using the same neural network architecture. The structure of each local discriminator is shown in Figure 4. The network features three middle layers with Layer Normalization (LayerNorm), followed by a fully connected layer that outputs a score. The parameters of the fully connected layer are determined based on the size of the cropped image.

The generated face $\hat{I}^b$ is cropped and fed into the corresponding local discriminators, producing a score list for each part. The scores in the list are then used to calculate the loss function. The specific cropping dimensions depend on the size of the generated image. The calculation process is detailed in Figure 13.

Equation (4) is the calculation formula for the GP penalty term, where $\tilde{x}$ is the generated image and x is the real image.

$$GP\_Loss = \underset{\tilde{x}\backsim P_g}{E}[D(x)] - \underset{x\backsim P_r}{E} + [D(x)] + \lambda \underset{\tilde{x}\backsim P_{\tilde{x}}}{E}[(||\nabla_{\tilde{x}} D(\tilde{x})|| - 1)^2] \tag{4}$$

In D1, there are D_{face}, D_{eyes}, D_{nose}, and D_{mouth}. In this study, the face, eyes, nose, and mouth are cropped to sizes of 85×85, 41×101, 35×29, and 21×41, with dimensions H×W. After inputting a real image X into D1, a segmentation function crops the image into these four parts, which are then fed into

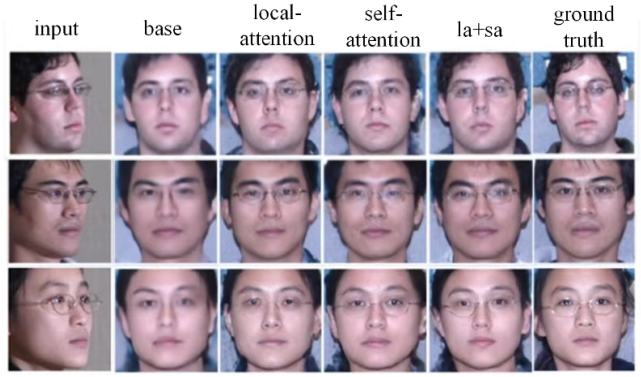

FIGURE 11. Different structures (attention mechanisms) generate results

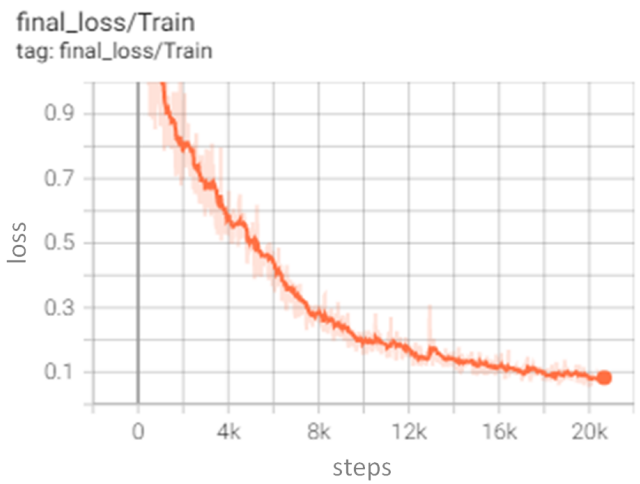

FIGURE 12. Variation trend of generator loss

the corresponding local discriminators. Each discriminator, following the network shown in Figure 4, outputs a scalar value, or score. These four scalar scores form a score list s^t. The average of these four scores is computed to obtain s^1. Similarly, the average score for the generated images is s^2. By applying the first part of Equation (4) to s^1 and s^2, the conventional adversarial loss is determined. The second part of Equation (4) is used to process the local positions separately to obtain the gradient penalty term. These two components are combined to yield the final *GP_Loss*, which is used to update the parameters of the improved D1 discriminator. The D2 discriminator retains its original structure, but the training method also employs the gradient penalty (GP) strategy to stabilize the training process.

Figure 6 illustrates the structure of the D2 discriminator, which uses target pose embeddings as conditions. It pairs target faces or generated faces as inputs to capture local perceptual information. The input to the D2 discriminator consists of a channel-wise overlay of frontal images and keypoint heatmaps. This input is processed through a series of convolutional, normalization, and activation layers, culminating in a sigmoid layer that produces a 6×6 single-

The process of the local discriminator

Parameter: s^t, s^f: The list of local output scores

1: $s^t \leftarrow$ D1(x)//s^t: Real output，x: true face

2: $s^1 \leftarrow \text{mean}(s^t)$ // Take the average score of the output

3: $s^f \leftarrow$ D1($\hat{x}$)// s^f: Generated output，$\hat{x}$:fake face

4: $s^2 \leftarrow \text{mean}(s^f)$ // Take the average score of the output

5: GP_loss $\leftarrow \text{GP}(s^1, s^2)$ // Gradient penalty GP loss

6: Update D1 parameters

FIGURE 13. local Discriminator train

Positive face training process

Parameter：D: Multi-angle face dataset；F: Frontal face dataset；E:epochs；G: Generator；D1: Discriminator1；D2: Discriminator2；L_r: Real landmark image；L_f: Frontal landmark image；θ: Model parameters

1: for t=0, . . . , E do

2: Sample a batch of profile face data X from the training set D as the current training data. Extract the corresponding frontal face of the same identity from F as the target generated sample Y, and obtain the corresponding facial keypoint landmark maps for both. L_r, L_f;

3: $X_{syn} \leftarrow G(X, L_r, L_f)$; // X_{syn}: Generated frontal face

4: score $\leftarrow$ D1(X_{syn},Y); // score: Scores of the generated faces from D1 compared to the real faces

5: $\theta_{D1} \leftarrow Adam(\theta_{D1};\ score)$ // Use the Adam optimizer to update the parameters of D1, adding gradient penalty (GP).

6: map $\leftarrow$ D2($(X_{syn}, L_f), (Y, L_f)$); // map: Obtain the probability map from D2.

7: $\theta_{D2} \leftarrow Adam(\theta_{D2};\ map)$; // Update the parameters of the D2 discriminator using gradient penalty (GP) based on the map.

8: total_loss $\leftarrow$ (G; D1; D2); // Calculate the total loss based on the outputs from G, D1, and D2.

9: $\theta_G \leftarrow Adam(\theta_G;\ total\_loss)$; // Update the parameters of the generator G using the total_loss.

10: end for

FIGURE 14. Model training algorithm

channel probability map. Each position in the probability map corresponds to a local region, capturing local perceptual information effectively.

D. LOSS FUNCTION

We use a multi-scale pixel-level L1 loss to constrain content consistency, formulated as follows:

$$L_{pix} = \frac{1}{S}\sum_{S=1}^{3}\frac{1}{W_s H_s C}\sum_{w,h,c=1}^{W_s,H_s,C}\left|\hat{I}^b_{s,w,h,c} - I^b_{s,w,h,c}\right| \tag{5}$$

TABLE 1. Rank-1 recognition rates on Multi-PIE dataset under Setting1

Method	±75°	±60°	±45°	±30°	±15°	Avg
TP-GAN [18]	77.43	87.72	95.38	98.76	98.80	91.45
FF-GAN [20]	77.20	85.20	89.70	92.50	94.60	87.85
PIM1 [29]	**92.50**	96.60	98.60	99.30	99.40	**97.28**
PIM2 [29]	91.20	**97.40**	98.30	99.40	99.80	97.22
Our	88.26	94.30	**98.97**	**99.70**	**99.87**	96.84

TABLE 2. Rank-1 recognition rates on Multi-PIE dataset under Setting2

Method	±75°	±60°	±45°	±30°	±15°	Avg
TP-GAN [18]	79.70	87.72	95.38	98.76	98.80	92.07
FF-GAN [20]	80.20	86.35	90.70	94.27	96.59	89.62
PIM1 [29]	**92.60**	96.72	98.63	99.41	99.50	97.37
PIM2 [29]	91.75	**97.50**	98.40	99.50	99.83	97.39
Our	91.00	97.20	**99.24**	**99.72**	**99.91**	**97.41**

S consists of 3 scales. In this study, feature maps are fused at dimensions 32×32, 64×64, and 128×128. L1 pixel loss is calculated separately for each of these scales. To integrate prior knowledge of data distribution, reduce the smoothness of synthetic images, and apply local attention, D_1conditional adversarial loss is used in the discriminator. The loss for distinguishing local features of synthetic images is as follows:

$$\begin{aligned} L^{D_1}_{adv} = E\left[D_1^{face}\left(I^b, \hat{I}^b\right) + D_1^{eye}\left(I^b, \hat{I}^b\right)\right] + \\ E\left[D_1^{nose}\left(I^b, \hat{I}^b\right) + D_1^{mouth}\left(I^b, \hat{I}^b\right)\right] \end{aligned} \tag{6}$$

Each of the four local feature regions is discriminated separately, and the parameters are updated based on the scores. The loss for evaluating local structural reconstruction information is as follows:

$$L^{D_2}_{adv} = E\left[\log D_2\left(I^b, P^b\right)\right] + E\left[\log\left(1 - D_2\left(\hat{I}^b, P^b\right)\right)\right] \tag{7}$$

This study uses the pre-trained Light-CNN29[23] to extract facial features while preserving facial identity attributes. The formula is as follows:

$$L_{id} = \left\|Net\left(\hat{I}^b\right) - Net\left(I^b\right)\right\|_2^2 \tag{8}$$

Net represents the identity extraction network. $\hat{I}^b$ and I^b denote the synthesized frontal face and the real frontal face, respectively. $\|\cdot\|_2$ is the L2 norm of their difference. Total Variation (TV) regularization loss is responsible for removing artifacts from the synthetic image, where C, H, W represent the number of channels, image height, and width, respectively.

$$L_{tv} = \sum_{c=1}^{C}\sum_{w,h=1}^{W,H}\left|\hat{I}^b_{w+1,h,c} - \hat{I}^b_{w,h,c}\right| + \left|\hat{I}^b_{w,h+1,c} - \hat{I}^b_{w,h,c}\right| \tag{9}$$

The overall loss is a weighted sum of the aforementioned losses, with different loss weights controlled by their respective coefficients.

$$\begin{aligned} \min_G \max_{D_1 D_2} L = \lambda_1 L_{pix} + \lambda_2 L^{D_1}_{adv} + \lambda_3 L^{D_2}_{adv} + \\ \lambda_4 L_{id} + \lambda_5 L_{tv} \end{aligned} \tag{10}$$

IV. EXPERIMENTS AND ANALYSIS

A. EXPERIMENTAL SETUP

This study conducts experiments on two datasets. The first is the widely used Multi-PIE [24] facial dataset, which includes faces of 337 individuals captured from different angles and

TABLE 3. Rank-1 recognition rates of different methods on CAS-PEAL-R1 dataset

Yaw Angle	0°	±15°	±30°	±45°	Avg	±15°	±30°	±45°	Avg	0	±15°	±30°	±45°	Avg
TP-GAN [18]	98.86	98.94	98.89	97.62	98.58	**100.0**	99.94	98.71	99.95	97.68	97.73	97.45	95.83	97.17
CR-GAN [26]	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 [30]	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
Our	**99.66**	**99.70**	**99.58**	96.47	98.85	**100.00**	**100.00**	**99.70**	**99.90**	**99.77**	**99.58**	**99.24**	97.35	**98.99**

TABLE 4. Rank-1 recognition rate under different changes on Multi-PIE

Method	±75°	±60°	±45°	±30°	±15°	Avg
Baseline	82.15	86.63	93.70	97.30	99.70	91.89
Local Attention	87.43	94.80	98.50	99.57	99.96	96.05
Self-Attention	86.35	93.88	97.85	99.21	99.92	95.44
Composite Attention	**90.26**	**95.30**	**98.97**	**99.70**	**99.98**	**96.84**

under various lighting conditions in a controlled indoor environment. In Setting 1, 250 individuals are used, with the first 150 subjects' faces for training and the remaining 100 for testing. In Setting 2, all 337 individuals are involved, with the first 200 used for training and the remaining 137 for testing. The dataset includes samples at 11 angles ranging from -75° to +75°. The second dataset is the CAS-PEAL-R1 [25], an Asian facial dataset collected by the Chinese Academy of Sciences. It contains various poses and lighting conditions for 1,040 subjects. This dataset includes samples of both yaw and pitch angles, enhancing the diversity of poses and training samples. The first 500 subjects are used for the training set, with the remaining 300 as the test set, ensuring no overlap.

The training method follows the standard GAN approach, utilizing paired data for training, which includes a frontal face and the corresponding profile face of the same identity I^a, I^b . Let I^a represent the frontal face and I^b represent the profile face. The generated frontal face $\hat{I}^b$ and the real frontal face I^b are both fed into the discriminator for evaluation. The training procedure is outlined in Figure 14. All input face images are aligned and cropped to a size of 128x128 pixels. Since the faces in the CAS-PEAL-R1 dataset are grayscale images, they are converted to RGB three-channel images using image channel operations before being fed into the network. The network model is implemented using the PyTorch deep learning framework. The Adam optimizer is selected for training, with the hyperparameters set as follows: learning rate = 2e-4, 1 = 0.5, and 2 = 0.999. The training process spans 50 epochs. The weights of the loss function are set to: 1 = 10, 2 = 0.1, 3 = 0.1, 4 = 0.02, and 5 = 1e-4.

B. EXPERIMENTAL ANALYSIS

This section presents various frontal faces generated using the improved model, trained and tested on the Multi-PIE and CAS-PEAL-R1 datasets. For the Multi-PIE dataset, the frontalization results for yaw angles are provided, while the CAS-PEAL-R1 dataset also includes results for pitch angles. Frontalization results are shown for faces with different angles, identities, and expressions. The evaluation metric used is the Rank-1 recognition rate, demonstrating the effectiveness of this frontalization method both qualitatively and quantitatively.

Results in Figure 7 show that after training for a certain number of epochs, side faces can generate frontal faces that preserve identity. In the early training stages, there may be artificial artifacts and unrealistic traces in certain facial areas. For instance, the eye regions appear blurry in the first and third rows of Figure 7(a). However, this issue improves as training progresses. Figure 8 demonstrates frontalization results for the same identity at different angles. The first row shows four angles: -60°, -45°, -30°, and -15° from left to right. It can be observed that even with large-angle deviations, a frontal face view can be restored. Figure 9 shows the frontalization results of faces with a smiling expression at four different angles. It can be seen that the model effectively corrects and maintains facial expressions in the frontal face. Figure 10 presents the frontalization results for faces with different pitch angles, in addition to yaw angles. Unlike the Multi-PIE dataset, which only includes yaw variations, this method successfully restores faces from the CAS-PEAL-R1 dataset to a standard frontal view, even from elevated angles. This demonstrates the feasibility of frontalizing faces with various poses and angles. This section tests the ability to maintain identity after face frontalization, a key goal being the consistency of identity between frontal and side views. Pre-trained LightCNN is used to extract facial feature vectors for comparison. As shown in Table 3, the proposed method outperforms others at angles from +15° to ±45°. Although PIM1 achieves the highest accuracy of 92.50% at +75°, our method's average accuracy is slightly lower than PIM due to the impact of larger angle deviations. However, overall performance is comparable to state-of-the-art methods. Analysis of Table 4 indicates that under the setting-2 configuration, our method performs exceptionally well, achieving an average recognition rate of 97.41%. This improved performance is attributed to the use of more training data in setting-2, enhancing the model's capabilities. Table 5 presents the Rank-1 recognition rates of different models on the CAS-PEAL-R1 dataset. Our method achieves the highest average accuracy for frontalization across different pitch angles.

C. ABLATION STUDY

In Figure 11, the input is a left-turned face at 60°. From left to right, the columns represent the results with different attention mechanisms added. The second column is the baseline model, the third column shows results with the discriminator incorporating local attention, the fourth column adds only self-attention, and the fifth column includes both attention layers. The rightmost column is the ground truth. Table 6 lists

the Rank-1 face recognition rates for different models. In this study, methods with added attention mechanisms required only 40 epochs to complete training, whereas those without required 50, indicating that both attention mechanisms enhance learning speed. Figure 12 shows that the generator's total loss confirms stable convergence of the model.

V. CONCLUSION

This paper proposes a generative adversarial network combining frontal pose keypoint guidance with attention mechanisms for generating frontal faces. The generator uses a U-net encoder-decoder structure to integrate multi-scale facial features. Input side-face images are overlaid with frontal pose keypoints on the channel layer and fed into the generator for feature extraction. To capture the global dependencies of feature maps, a self-attention module is added to the decoder, enhancing the network's adaptability to lighting and other image information. Two discriminators are used, one employing local attention for supervised learning. The result is identity-preserving frontal faces suitable for facial recognition and other tasks like face editing and dataset expansion. Qualitative and quantitative experiments confirm that combining pose keypoints with attention mechanisms improves the quality of generated images.

REFERENCES

[1] K. Cao, Y. Rong, C. Li, X. Tang, and C. C. Loy, "Pose-robust face recognition via deep residual equivariant mapping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5187–5196, 2018.

[2] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker, "Reconstruction-based disentanglement for pose-invariant face recognition," in *Proceedings of the IEEE international conference on computer vision*, pp. 1623–1632, 2017.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[4] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1, pp. 786–791, IEEE, 2005.

[5] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4295–4304, 2015.

[6] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 787–796, 2015.

[7] G. Passalis, P. Perakis, T. Theoharis, and I. A. Kakadiaris, "Using facial symmetry to handle pose variations in real-world 3d face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1938–1951, 2011.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[9] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8398–8406, 2018.

[10] M. Mirza, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[11] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, pp. 214–223, PMLR, 2017.

[12] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

[13] 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 *Proceedings of the IEEE international conference on computer vision*, pp. 2439–2448, 2017.

[14] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, *et al.*, "Towards pose invariant face recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2207–2216, 2018.

[15] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proceedings of the IEEE international conference on computer vision*, pp. 3990–3999, 2017.

[16] V. Mnih, N. Heess, A. Graves, *et al.*, "Recurrent models of visual attention," *Advances in neural information processing systems*, vol. 27, 2014.

[17] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*, pp. 7354–7363, PMLR, 2019.

[18] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*, pp. 7354–7363, PMLR, 2019.

[19] R. Abdal, Y. Qin, and P. Wonka, "Image2stylegan: How to embed images into the stylegan latent space?," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4432–4441, 2019.

[20] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[21] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas, "Cr-gan: learning complete representations for multi-view generation," *arXiv preprint arXiv:1806.11191*, 2018.

[22] Y. Qian, W. Deng, and J. Hu, "Unsupervised face normalization with extreme pose and expression in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9851–9858, 2019.

[23] 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.

[24] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and vision computing*, vol. 28, no. 5, pp. 807–813, 2010.

[25] 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, 2007.

[26] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, *et al.*, "Towards pose invariant face recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2207–2216, 2018.

[27] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10043–10051, 2019.

[28] 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.

[29] Y. Qian, W. Deng, and J. Hu, "Unsupervised face normalization with extreme pose and expression in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9851–9858, 2019.

[30] S. Afra and R. Alhajj, "Face reconstruction from profile to frontal evaluation of face recognition," *Data Management and Analysis: Case Studies in Education, Healthcare and Beyond*, pp. 117–133, 2020.

JIHUI ZHAO is currently working toward the M.S. degree at Qingdao University of Science and Technology. His research interests include deep learning and Artificial Intelligence.

ZHENGYI YUAN received the B.S. degree in computer science and technology from the Qingdao University of Science and Technology, China, in 2017. He is currently working toward the M.S. degree at Qingdao Uni-versity of Science and Technology. His research interests include deep learning and computer vision.

JUNHU ZHANG received the Ph.D. degree in computer software and theory from the Department of Computer Science, Peking University, Beijing, China, in 2006. From 2006 to 2008, he was a Postdoctoral Fellow with the Sino-French Joint Laboratory, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently an Associate Professor and Master's Supervisor with the Department of Computer Science, where he focuses on research in AI for computer vision and AI for numerical modeling. Dr. Zhang's research interests include artificial intelligence, computer vision, and numerical modeling techniques. Technology. His research interests include deep learning and computer vision.

HAITAO LI received the Ph.D. degree from Ocean University of China. He is currently an Associate Professor and Master's Supervisor at Qingdao University of Science and Technology. He has been recognized as a Leading Talent in Qingdao City, an Innovation and Entrepreneurship Leading Talent in Qingdao, and a "5313" Science and Technology Entrepreneurship Leading Talent in Zhoushan City. Dr. Li is also a national leader in smart fisheries and serves as Vice President of the China Fisheries Association. Additionally, he is the Director of the Shandong Province Smart Ocean and Aquaculture Big Data Collaborative Innovation Center and the Secretary-General of the Qingdao Ocean Information Service Industry Alliance.

HUI LI received the Ph.D. degree in Computer Application Technology from Wuhan University of Technology, Wuhan, China, in 2013. He also obtained his Master's degree in Computer Application Technology from Wuhan University of Technology in 2010 and his Bachelor's degree in Computer Science and Technology from Henan Polytechnic University, Jiaozuo, China, in 2007. His research interests include computer vision, visual perception technology for autonomous driving, deep learning for 3D multi-object detection and tracking, and trajectory prediction.