International Core Journal of Engineering Volume 10 Issue 1, 2024
ISSN: 2414-1895 DOI: 10.6919/ICJE.202401_10(1).0021

A Multi-pose Face Frontalization Reconstruction Method for Uncontrolled Scenes

Haoyuan Guo

College of Computer Science and Technology, Taiyuan Normal University, Shanxi 030619, China

Abstract

Face recognition research in extreme posture and extreme environment has strong practical significance. Aiming at the problem of low face recognition accuracy caused by different face poses, face frontalization reconstruction provides an effective method. In this regard, this paper proposes a new GAN-based network model. In the generator, a multi-scale feature fusion recurrent neural network is included, which can effectively capture the hierarchical structure and context information in face data, so as to generate better face images that can maintain identity. In the discriminator, a new face attention mechanism is integrated to focus on the important local features of the face region, thereby enhancing the realism of the synthesized face. In this paper, training and verification are performed on Multi-PIE and CAS _ PEAL _ R1 datasets, and verification is performed on the uncontrolled dataset LFW. The method proposed in this paper has better face details in the objective index Rank-1 recognition rate to 95.37 %, and the visual quality. The quantitative and qualitative comparison is better than the comparison method.

Keywords

Multiple Poses; Face Frontalization; GAN; Multi-scale Recursive Network; Attention Mechanism.

1. Introduction

Automatic face recognition from images and video frames has always been a hot research topic. At present, in some very challenging data sets [1,2], data-driven models have achieved good recognition results. These face recognition technologies are often used in scenarios such as authentication of mobile devices. However, these scenes are carried out with the knowledge of the subject, and the face posture, expression, illumination and so on are ideal. In uncontrolled scenes such as video surveillance, the face recognition rate is often affected by illumination and pose problems. For example, the pre-training models VGGFace [3] and LightCNN [4] with better face recognition effect have only 2.1 % and 5.5 % accuracy when recognizing extreme poses with a yaw angle of 90 °. Face recognition research in extreme posture and extreme environment has strong practical significance.

The mainstream methods for multi-pose face recognition can be categorized into two groups.The first is to extract pose-robust features directly from the original image. This method can extract very limited features due to pose constraints when dealing with extreme pose faces, resulting in a significant reduction in face recognition performance. The second method is to first frontalize the face, and then extract features. This method can extract more effective discriminant features, thereby improving the performance of face recognition. Facial frontalization aims to synthesize a frontal view using a given contour. The composite image can be directly used for general face recognition methods without the need to specify additional complex modules. In addition to facial recognition, creating

ISSN: 2414-1895 DOI: 10.6919/ICJE.202401_10(1).0021

realistic frontal faces is also conducive to a range of face-related tasks such as face reconstruction, face attribute analysis, and face animation.

Using image generation technology, Xu et al. (2019) [5] proposed a multi-task convolutional encoder-decoder network (MCEDN). This approach introduces a frontal basic feature network to generate the fundamental characteristics of the frontal facial region, and furthermore, integrates local characteristics of the multi-pose face extracted by the coding network to enhance the overall representation. Yu et al. (2020) [6] proposed a Transform Discriminant Neural Network (MTDN), which can simultaneously achieve front image creation and upscaling. Although the above methods have achieved good results, the synthesized faces lack exquisite details and are often blurred and untrue in large poses. Since Goodfellow et al. [7] proposed the generative adversarial network, GAN is extensively employed in the realm of image generation. In recent years, many face frontalization reconstruction methods based on GAN have achieved remarkable results (Cao et al., 2019 [8] ; zhou et al., 2020 [9] ; tu et al., 2022 [10]). Cao et al.proposed a high-fidelity pose invariant model (HF-PIM) to obtain high-resolution realistic and identity-preserving results. HF-PIM combines the advantages of methods based on 3D and GAN, and performs frontal processing on contour images through a new facial texture fusion distortion program. Zhou et al.proposed a new unsupervised framework to solve the limitation of the size and scope of data sources. They constructed self-supervision by repeatedly rotating and rendering 3D face modeling, and used ordinary CycleGAN to generate the final image. The framework does not need to rely on multi-view images of the same person, and can generate high-quality face images from other perspectives. However, 3D-based methods often require more parameters and calculations. Tu et al.proposed a face pose normalization module, which guides the face frontalization reconstruction process by analyzing the gap between the input face and the reference face pose. But the generated portrait has accidental artifacts. Inspired by [11, 12], this method integrates multi-scale recurrent neural network into the generator [13].Combined with multi-view prediction and feature fusion, the generated frontal face image has more identity features of the original image, and the face attention mechanism is added to the discriminator [14], so that the generated portrait has more fine texture.

In this paper, the generator G and discriminator D of the GAN network are redesigned, and the training is countered in a complementary way. On the famous Multi-PIE face recognition benchmark, the experimental effect can be improved to 95.37 % on the objective index Rank-1 recognition rate, and the superiority of this method is verified by comparative experiments.

2. GAN Network Combining Multi-scale Recursion and Face Attention

2.1 Face Frontalization

Suppose that P_{data} is a data set containing frontal and lateral facial images. Let $\{I^f, I^p\}$ be a pair of frontal and lateral face images of the same person sampled from P_{data}. Given a side face imageI^p, the purpose of the experiment is to train the generator G to synthesize the corresponding front face image $\hat{I}^f$=$G(I^p)$, and the generated front face image can maintain the identity of the person and visually approach I^f.

In order to achieve this, this paper proposes the model structure shown in Figure 1 to train the target generator G. The model has two main components, namely, a multi-scale recurrent neural network G and a facial attention discriminator D. G can handle the features of different levels and scales of input data, so it has better feature representation. At the same time, the face attention in D is identified by four separate discriminator models for different features of the face. Thus, the local consistency of I^p and I^f is enhanced. In this way, the model can generate more realistic frontal images and maintain the identity of the face.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202401_10(1).0021

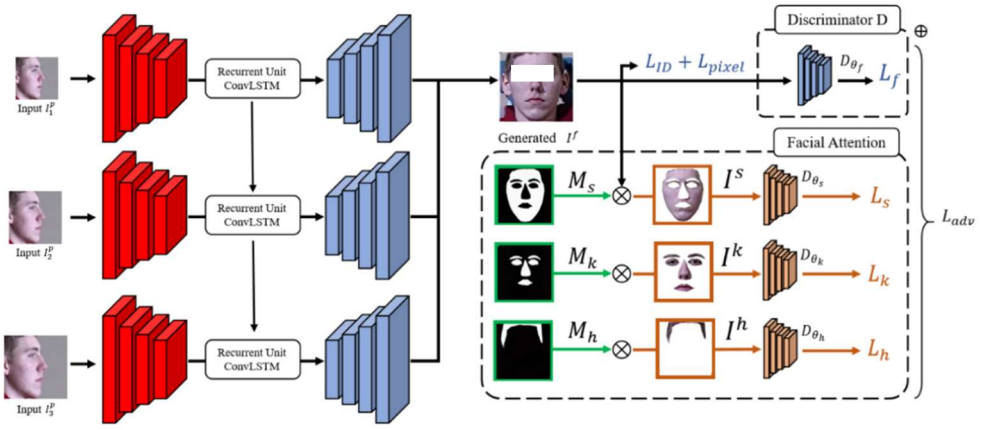

Fig. 1 Network structural diagram

2.2 Multiscale Recurrent Neural Network.

This network structure is mainly used to generate frontal face images. This part combines multi-scale recurrent neural networks with feature fusion to generate pixel-level predictions. UNet / DeepLabV3 + [15] is used as the basic network, and its function is to complete feature encoding and decoding. The architecture is a three-level encoder-decoder segmented network with recursive units. Each level comprises four convolutional layers adorned with ReLU activation, a ConvLSTM module [16] and three deconvolution layers. There are three Resnet blocks behind each layer above. To regain the missing detailed spatial data, the replication connection links the features of the convolutional layer with the features of the deconvolution layer through matrix addition. The tail consists of a convolutional layer coupled with a Sigmoid activation function, enabling pixel-level classification.

LSTM correlation algorithm is usually used to process time series data. In order to be more effective in image feature extraction, we introduce ConvLSTM into the multi-level generator network to achieve global feature transfer. The main function of ConvLSTM is to pass long-term and short-term memory to the next level. By using ConvLSTM, we can optimize the encoded feature maps from different levels. Considering the three-tier architecture of the network, ConvLSTM can enhance feature information from coarse to fine. In this architecture, the unit state C_l and the hidden state H_l are responsible for long-term memory and short-term memory, respectively, where l represents the level index.

$$C_l, H_l = g(C_{l-1}, H_{l-1}, E_l) \tag{1}$$

where l represents the feature of the encoder from level l. In the first level, ConvLSTM input is insufficient due to the lack of long-term or short-term memory from the previous level. Therefore, in the initialization, the unit state C_0 and the hidden state H_0 are replaced by a zero matrix. Since the short-term memory is suitable for running at the same level, the hidden state H_l is transmitted to the decoder at the corresponding level.

By amassing the above levels of generated graphs, the third-level network of this architecture will produce the final result. The third-level feature fusion algorithm makes the result better than individual forecast. The size of the output feature R_s is adjusted to the size of the input image, and recursively fused :

$$R_s = u\big(p(F_{s-1})\big) + p(F_s) \tag{2}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202401_10(1).0021

Among them, s is the s-th scale in the third level, F_s is the corresponding decoder mapping, u(*) increases the sample * according to the ratio of 2, and p(F_s) is the prediction graph in the s-th scale.

The loss of this network architecture is the sum of three levels, and each level optimizes its own parameters at the same time. The input image sizes are (w/4, h/4) , (w/2, h/2)and (w, h). In the first two levels, the input X_l and the true value Y_l reduced in size by the weighted average of pixels in a 2-by-2 neighboring area. l represents the level index. Then, the L_{SB} loss uses the Dice function d to assess the difference between the generated output f (·, θ) and the true value :

$$L_{SB} = \sum_{l=1}^{3}(1 - d(f(X_l, \theta), Y_l) \tag{3}$$

$$d(f(X_l, \theta), Y_l) = 2 \times \left[\frac{\sum_{i=1}^{p}(f(X_l,\theta)_i) \times Y_{li})}{\sum_{i=1}^{p}(f(X_l,\theta)_i) + Y_{li})}\right] \tag{4}$$

where p is the number of pixels of the original image, and θ is the weight.

2.3 Face Attention Mechanism

In order to synthesize realistic frontal face photos, the generation model must focus on each face detail, rather than distinguish the whole face. Therefore, a new face-attention scheme is further introduced, which guides the discriminator by using three additional segmentations, which cooperate with the discriminator D_f but concentrate on various facial areas. Specifically, the front is divided into three local areas (skin, key points, and hairline), and each area is assigned to a region discriminator (D_s, D_k, and D_h). Each region discriminator tends to improve the composite images in their respective regions and complement each other.

Inspired by [17], the frontal face is parsed into three predefined regions. Specifically, the pre-trained model [18] is used as a ready-made face parser f_P to generate three masks, and then they are applied to frontal face images to create regional images. These images are low-frequency regions I^s (i.e., skin regions), key point features I^k (i.e., eyes, eyebrows, nose and lips) and hairline I^h. Mathematically speaking.

$$M_s, M_k, M_h = f_P(I^f) \tag{5}$$

Among them, M_s, M_k and M_h do masks cover skin, key feature points, and hairline regions. The subscripts align with the characteristics. Therefore:

$$\begin{aligned} fake \quad & I^s = I^f \odot M_s, I^k = I^f \odot M_k, I^h = I^f \odot M_h; \\ real \quad & \hat{I}^s = I^f \odot M_s, \hat{I}^k = I^f \odot M_k, \hat{I}^h = I^f \odot M_h. \end{aligned} \tag{6}$$

where $\odot$ is the product of elements, and $\hat{I}^s$, $\hat{I}^k$ and $\hat{I}^h$ represent the regional images of skin, key points and hairline.

The four discriminators (D_f, D_s, D_k and D_h) distinguish the real frontal images of the four views (I^f, I^s, I^k and I^h) and the corresponding composite frontal images after their superscripts ($\hat{I}^f$, $\hat{I}^s$, $\hat{I}^k$ and $\hat{I}^h$). Each discriminator D is trained in an adversarial manner using the generator G. Thus, facial attention loss is composed of four independent discriminators with adversarial loss.

$$L_j = E_{I^j}[logD_f(I^j) + E_{\hat{I}^j}[log(1 - D_j(\hat{I}^j))] \tag{7}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202401_10(1).0021

where j ∈ { f, s, k, h }. Each D_j maximizes its objective function L_j, while G minimizes it. The complete target can be represented by the minimum-maximum formula:

$$\min_G \max_D L_{adv}(D, G) \tag{8}$$

L_{adv} is the total loss of adversarialness.

$$L_{adv} = \sum_{j\in\{f,s,k,h\}} L_j(D_j, G) = \sum_{j\in\{f,s,k,h\}} (E_{Ij}[logD_j(I^j)] + E_{\hat{I}j}[log(1 - D_j(\hat{I}^j))]) \tag{9}$$

Where j ∈ { f, s, k, h }, the losses are L_f, L_s, L_k and L_h. Each region is focused on optimizing its composite image and complementing each other.

2.4 The Object Function of G

(1) Loss of identity retention : A key aspect of positive evaluation is to retain identity in the process of positive synthesis. The pre-trained face recognition network is used to extract key features to improve G' s identity retention ability. Specifically, we use a pre-trained 29-layer LightCNN [4], whose weights are fixed during training, to calculate G' s identity protection loss. The identity protection loss is defined as the feature level difference in the last two fully connected layers of LightCNN between the composite front and the real front :

$$L_{ID} = \sum_{i=1}^{2} \left\| p_i(I^f) - p_i(\hat{I}^f) \right\|_2^2 \tag{10}$$

where pi (·) (i ∈ 1,2) is the output feature of the fully connected layer from Light CNN, and $||\cdot||_2$ is the L2 norm.

(2) Multi-scale pixel-by-pixel loss : Following [17], multi-scale pixel-by-pixel loss is used to constrain content consistency. The different layers of the decoder in G output multi-scale composite images. The loss of the i^{th} sample is the total mean difference between the multi-scale synthesis and the original front (i.e., $\hat{I}_i^f$ and I_i^f, respectively). Mathematically speaking :

$$L_{pixel} = \frac{1}{S}\sum_{s=1}^{S} \frac{1}{W_s H_s C} \sum_{w,h,c=1}^{W_s,H_s,C} \left| G\left(I_{s,w,h,c}^p\right) - I_{s,w,h,c}^f \right| \tag{11}$$

where S is the number of scales, W_s and H_s are the corresponding width and height of scale s. The resultant frontal $G\left(I_{s,w,h,c}^p\right) = \hat{I}_{s,w,h,c}^f$ is transformed by G with learning parameter θ_G. In this model, the parameters are set to S = 3, and the scales are 32 × 32,64 × 64 and 128 × 128.

(3) Total variation regularization : Total variation regularization L_{tv} is used to remove the artifacts in the synthetic image $\hat{I}^f$.

$$L_{tv} = \sum_{c=1}^{C} \sum_{w,h=1}^{W,H} \left| \hat{I}_{w+1,h,c}^f - \hat{I}_{w,h,c}^f \right| + \left| \hat{I}_{w,h+1,c}^f - \hat{I}_{w,h,c}^f \right| \tag{12}$$

where C, W and H represent the channel, width and height of $\hat{I}^f$.

(4) Total loss : The objective function is the weighted sum of the above losses : Here λ_1, λ_2, λ_3 and λ_4 are the hypter parameters that control the weight of the loss term.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202401_10(1).0021

$$L_G = \lambda_1 L_{ID} + \lambda_2 L_{pixel} + \lambda_3 L_{adv} + \lambda_4 L_{tv} \quad (13)$$

3. Experimental Results and Analysis

In order to illustrate that the model can ensure the fidelity of synthetic photos while maintaining identity features, quantitative and qualitative evaluations of the model were performed on controlled and field environmental data sets. In this section, we first introduce the selection of data sets and experimental details. Then, it is proved that the model in this paper is superior to the most advanced methods in terms of qualitative synthesis results and quantitative identification results. Finally, ablation studies were carried out to prove the advantages of each part of the model. Figure 2 shows the reconstruction results of the proposed method.

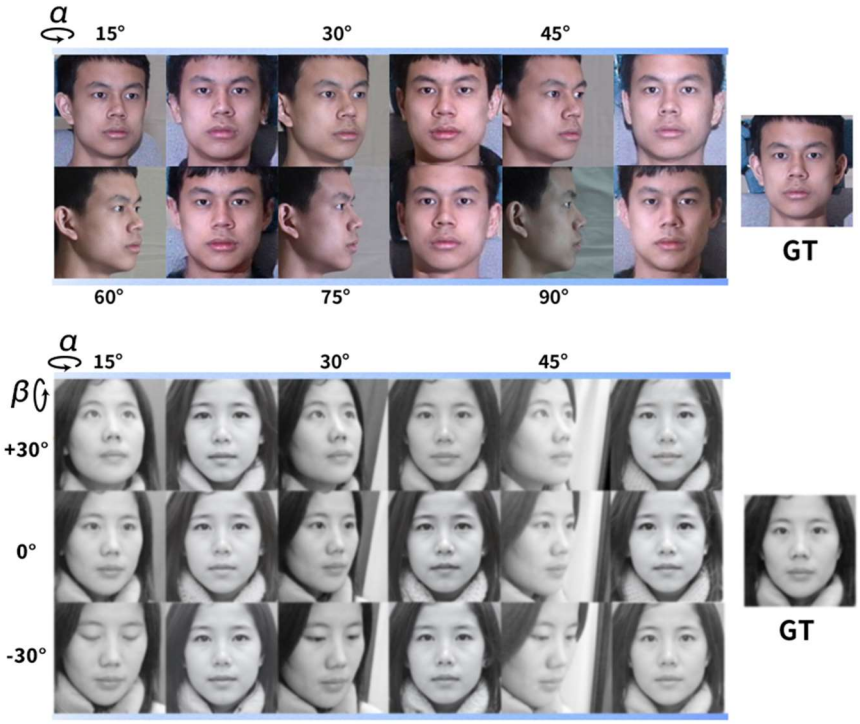

Fig. 2 Reconstruction results of this method

3.1 Dataset

The Multi-PIE dataset is considered to be the largest public dataset for face synthesis and recognition in a controlled environment. It includes 337 identities, involving 4 sessions (i.e., time), and contains 754,204 images from 15 countries. The experiment followed the second setting in [17,19] to emphasize the changes of posture, illumination and sessions. This setting includes 4 sessions and 337 identities of all neutral expression images. The front and side images of the first 200 subjects were used for training. The training samples included 13 postures with yaw angles in the range of ± 90 ° and 20 lighting conditions with different lighting conditions. The remaining 137 identity samples constitute the test set. This ensures that there is no overlap between the training set and the test set.

The CAS _ PEAL _ R1 dataset is a large public dataset of Chinese faces. Each identity has different postures, expressions, accessories, and illumination changes. A total of 30,863 grayscale images were obtained from 1,040 subjects (595 males and 445 females). Images of various attitudes are used, including 6 yaw angles (i.e., α = { 0 °, ± 15 °, ± 30 °, ± 45 ° }), 3 pitch angles (i.e., β = { 0 °, ± 30 ° }), a total of 21 yaw-pitch angle rotations. The first 600 subjects were used for training, and the remaining 440 subjects were tested.

LFW contains 13,233 face images collected in the wild environment (unconstrained environment). It will be used to evaluate model performance under uncontrolled settings.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202401_10(1).0021

3.2 Experimental Considerations Detail

In order to train the model, an image pair $\{ I^f, I^p \}$ consisting of a side view image and a corresponding front view image is required. After [17], we first cut all images into a standard view with a size of 128 × 128. For Multi-PIE, both the real image and the generated image are RGB images. For CAS-PEAL-R1, all images are grayscale. The identity preserving network used in the model is a pre-trained model on the MS-Celeb-1M dataset, but fine-tuned on the experimental dataset. The experimental parameters were set as follows : $\lambda_1 = 0.1$, $\lambda_2 = 10$, $\lambda_3 = 0.1$, $\lambda_4 = 1^{-4}$.

3.3 Face Synthesis

In this section, the synthesis results of the model in this paper are visually compared with the most advanced methods. Fig.3 shows the qualitative comparison of Multi-PIE, and shows the synthesis results of different methods [19,20,21] under extreme postures of 60 ° and 90 °. In this way, the superior performance of the method in large posture is proved. Qualitative results show that the frontal image restored by the proposed method has finer details (i.e., more accurate facial shape and texture), while the synthesis results of other methods lack accuracy. In order to show the authenticity of the image synthesized from any view, Fig.4 shows the synthesized front results of each pose from 0 to 90 °.

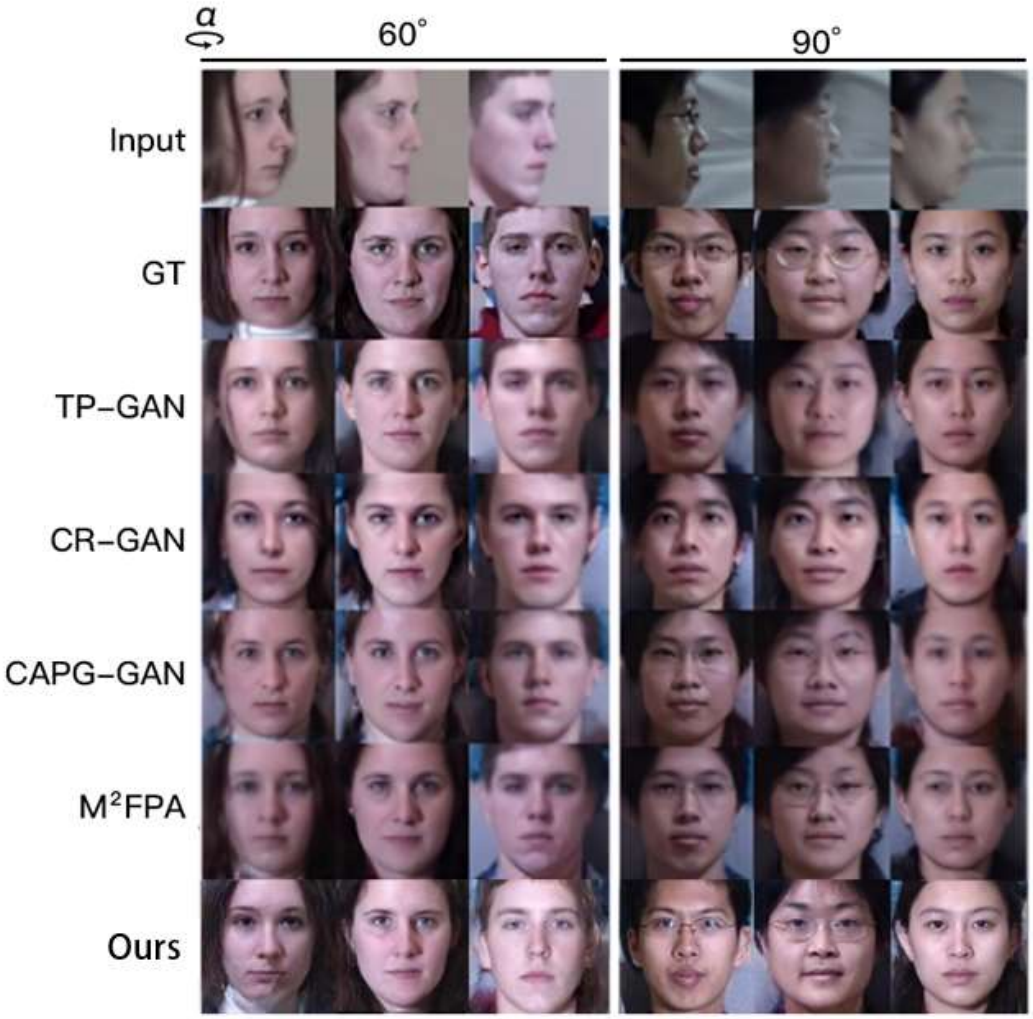

Fig. 3 The reconstruction results of different methods in Multi-PIE dataset

ISSN: 2414-1895 DOI: 10.6919/ICJE.202401_10(1).0021

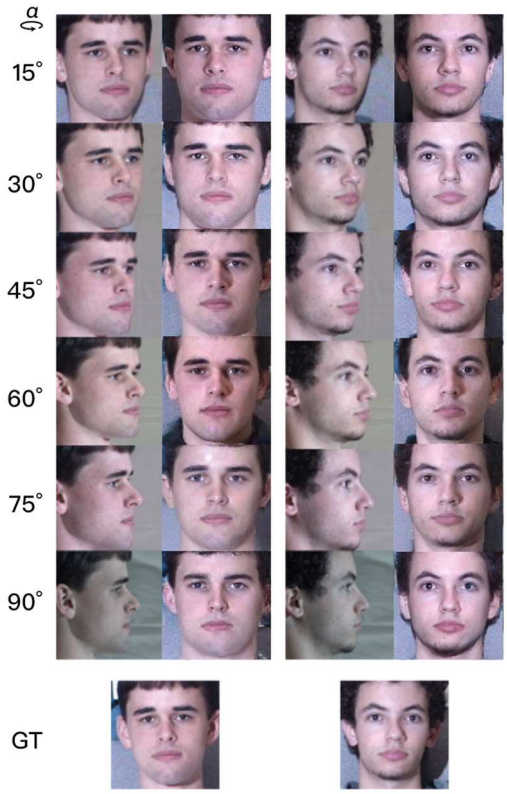

Fig. 4 Face reconstruction results from different angles of Multi-PIE dataset

In order to further verify the improved results of the model in multiple yaws and pitches, the results of the CAS-PEAL-R1 data set on different methods [19,20,21] are also compared, because it also contains the attitude change of the pitch angle. Fig. 5 shows that the proposed method can generate realistic faces (i.e., finer details) while retaining the identity of the person.

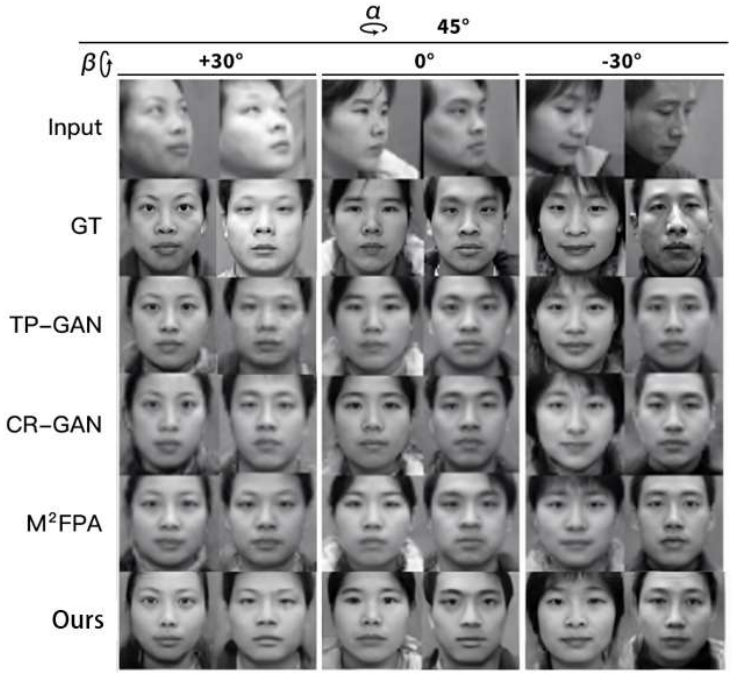

Fig. 5 The reconstruction results of different methods in CAS _ PEAL _ R1 dataset.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202401_10(1).0021

Fig. 6 The reconstruction results in the LFW dataset

3.4 Identity Protection Features

In order to quantitatively prove the identity retention ability of the proposed method, the face recognition accuracy of the synthesized frontal image is evaluated. Table I compares the performance of Multi-PIE in different postures with the existing excellent methods. The results were displayed with Rank-1 recognition rate. The experiment uses a pre-trained 29-layer LightCNN [4] as a face recognition model to extract features, and uses a cosine distance metric to calculate the similarity of these features. Larger poses often provide less information, making it more difficult to maintain the identity of the synthesis results. As shown in Table I, the performance of the existing methods decreases sharply as the attitude angle increases to 75 ° or even larger. The proposed method still has convincing performance in these extreme postures (i.e., 75 ° and 90 °). In addition, it can also achieve excellent results in other smaller postures (i.e., 15 ° ~ 60 °). Table II shows the Rank-1 recognition rate of CAS-PEAL-R1 in yaw (α) and pitch (β) attitude changes. The results also prove that the model has excellent identity protection ability in multiple poses.

The face verification performance (ACC and AUC) of the proposed method is compared with other advanced methods on the LFW benchmark, and the results are comparable. The quantitative results in Table III prove that the proposed method can effectively preserve identity information. The qualitative results of LFW are shown in Figure 6. For the qualitative results of LFW, some extreme angles are selected for testing, and it can still be seen that it has a good reconstruction effect.

Table 1. Rank-1 recognition rate comparison of multi-pose face reconstruction images in Multi-PIE dataset (%)

	±90°	±75°	±60°	±45°	±30°	±15°	Avg
LightCNN	5.51	24.18	62.09	92.13	97.38	98.59	63.31
TP-GAN	64.64	77.43	87.72	95.38	98.06	98.68	86.99
FF-GAN	61.20	77.20	85.20	89.70	92.50	94.60	83.40
CAPGGAN	66.05	83.05	90.63	97.33	99.56	99.82	89.41
PIM1	71.60	92.50	97.00	98.60	99.30	99.40	93.07
PIM2	75.00	91.20	97.70	98.30	99.40	99.80	93.57
M²FPA	75.33	88.74	96.18	99.53	99.78	99.96	93.25
BaseLine	66.08	84.21	90.84	97.71	99.25	99.70	89.63
G+MSRNN	77.14	89.78	96.12	99.21	99.78	99.99	93.67
D+face-attention	77.32	90.69	96.19	99.11	99.79	99.99	93.85
Ours	**81.78**	**93.49**	**97.91**	**99.11**	**99.91**	**99.99**	**95.37**

ISSN: 2414-1895 DOI: 10.6919/ICJE.202401_10(1).0021

Table 2. Rank-1 recognition rate comparison of multi-pose face reconstruction images in CAS-PEAL-R1 dataset (%)

	Pitch(-15°)					Pitch(0°)				Pitch(+15°)				
Yaw	±0°	±15°	±30°	±45°	Avg_1	±15°	±30°	±45°	Avg_2	±0°	±15°	±30°	±45°	Avg_3
TPGAN	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
CRGAN	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
Ours	**99.81**	**99.83**	**99.69**	**99.11**	**99.61**	**100.00**	**100.00**	**99.89**	**99.96**	**99.12**	**99.04**	**98.62**	**98.26**	**98.76**

Table 3. The results of face verification accuracy (ACC) and area under the curve (AUC) of LFW were analyzed.

	ACC(%)	AUC(%)
LFW-3D	93.62	88.36
LFW-HPEN	96.25	99.39
FF-GAN	96.42	99.45
CAPG-GAN	99.37	99.90
M²FPA	99.41	99.92
DA-GAN	99.56	99.91
Ours	**99.61**	**99.94**

3.5 Ablation Experiment

Through ablation experiments, the contribution of multi-scale recursive network in G and facial attention mechanism in D to frontalization reconstruction was analyzed. The baseline model consists of only a U-Net generator [15] and an ordinary frontal discriminator. The variants are constructed by adding a multi-scale recurrent neural network or a facial attention mechanism to the baseline model separately. In addition to the ablation study of the face attention mechanism, each discriminator used in the face attention scheme is also characterized.

(1) Effects of two variants : In order to highlight the importance of multi-scale recurrent neural networks in G and facial attention in D, Table I shows its quantitative results. The results show that the single use of any kind of optimization network will significantly improve the recognition performance, and the combination of the two will achieve the best performance, especially for large postures.

In addition, the comparison between the baseline model and the variant is also shown in the qualitative results (see Fig.7). The results show that the synthesis results with the addition of multi-scale recurrent neural networks have relatively less ambiguity (e.g., blurred faces and ears). However, from the perspective of visualization, it still has considerable ability to maintain identity. In contrast, models with only face attention can generate realistic faces, but retain less identity information. The network that combines multi-scale recursion and facial attention mechanism can generate a frontal view with relatively more details (i.e., facial appearance and texture).

ISSN: 2414-1895 DOI: 10.6919/ICJE.202401_10(1).0021

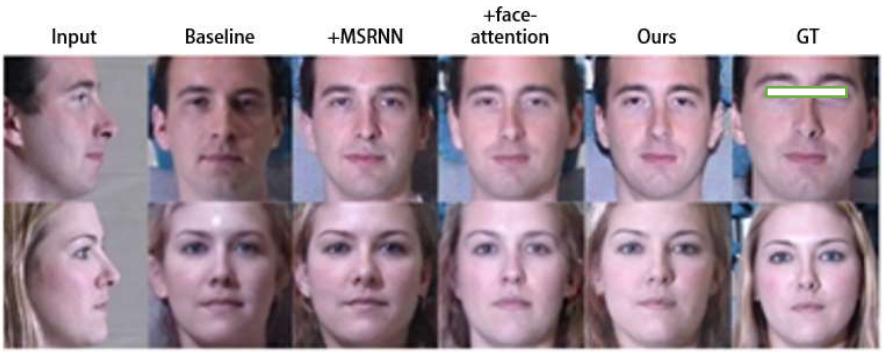

Fig. 7 Add the reconstruction results of different model components

(2) The effects of different masks used in D : Quantitative (Table IV) and qualitative (Figure 8) The results of three different masks used for facial attention in D were studied. The quantitative results show that the key point features have the most obvious influence on the face recognition task, and the hairline features have the least influence. In addition, a qualitative comparison between different mask variants is shown. By adding a hair discriminator (D_h), the model can generate relatively clear edges in the hair region. The skin discriminator (D_s) contributes to low-frequency features, while the key point discriminator (D_k) helps to generate real facial attributes (such as eyes) to determine real values. Finally, high-quality facial images are obtained by fusing all these independent discriminators.

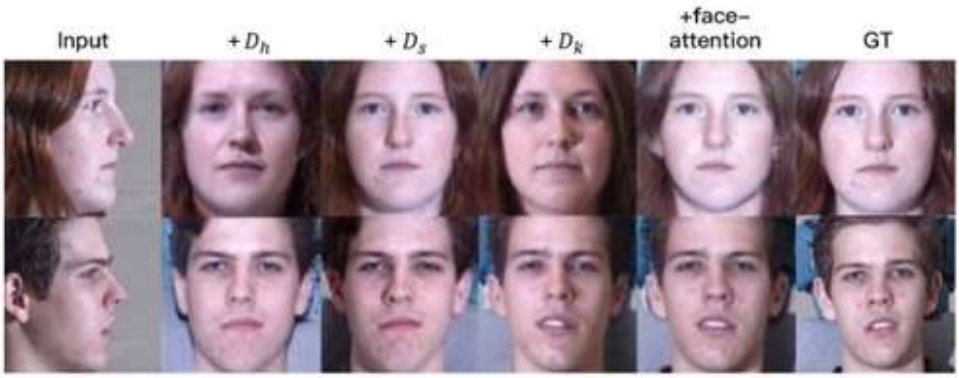

Fig. 8 Reconstruction results of adding different facial key point discriminators

Table 4. Ablation study : Rank-1 recognition rate of facial key points

	±90°	±75°	±60°	±45°	±30°	±15°	Avg
$D+D_h$	72.89	86.41	93.12	98.41	99.72	99.99	91.76
$D+D_s$	72.91	87.88	94.01	98.78	99.52	99.97	92.18
$D+D_k$	78.34	88.49	95.11	98.91	99.64	99.99	93.41
D+face-attention	**77.32**	**90.69**	**96.19**	**99.11**	**99.79**	**99.99**	**93.85**

4. Conclusion

In this paper, a multi-scale recurrent and facial attention mechanism fusion network is adopted. Based on the traditional GAN network, a multi-scale recurrent neural network is introduced in G, and then trained in an adversarial manner by D equipped with facial attention mechanism. During the experiment, the proposed network model effectively synthesizes the frontal view of the face in

ISSN: 2414-1895 DOI: 10.6919/ICJE.202401_10(1).0021

extreme poses. The experimental results show that the synthesized portrait can eliminate artifacts and make it more realistic. In extreme postures, it can still maintain identity features and improve the performance of face recognition. It is proved that the proposed method can effectively solve the problem of multi-pose face recognition.

References

[1] Cao Q, Shen L, Xie W, et al. Vggface2: A dataset for recognising faces across pose and age[C]//2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, 2018: 67-74.

[2] Yin Y, Robinson J, Zhang Y, et al. Joint super-resolution and alignment of tiny faces[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12693-12700.

[3] Parkhi O, Vedaldi A, Zisserman A. Deep face recognition[C]//BMVC 2015-Proceedings of the British Machine Vision Conference 2015. British Machine Vision Association, 2015.

[4] Wu X, He R, Sun Z, et al. A light CNN for deep face representation with noisy labels[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(11): 2884-2896.

[5] Haiyue Xu, Naiming Yao, Xiaolan Peng, et al. Codec network-based frontalization method for multi-gesture face images[J]. SCIENCE CHINA: Information Science, 2019.

[6] Yu X, Porikli F, Fernando B, et al. Hallucinating unaligned face images by multiscale transformative discriminative networks[J]. International Journal of Computer Vision, 2020, 128(2): 500-526.

[7] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[8] Cao J, Hu Y, Zhang H, et al. Towards high fidelity face frontalization in the wild[J]. International Journal of Computer Vision, 2020, 128: 1485-1504.

[9] Zhou H, Liu J, Liu Z, et al. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5911-5920.

[10] Tu X, Zhao J, Liu Q, et al. Joint face image restoration and frontalization for recognition[J]. IEEE Transactions on circuits and systems for video technology, 2021, 32(3): 1285-1298.

[11] Shen N, Xu T, Bian Z, et al. SCANet: A Unified Semi-supervised Learning Framework for Vessel Segmentation[J]. IEEE Transactions on Medical Imaging, 2022.

[12] Ma Z, Zhang H, Liu J. MS-RNN: A flexible multi-scale framework for spatiotemporal predictive learning[J]. arXiv preprint arXiv:2206.03010, 2022.

[13] Tao X, Gao H, Shen X, et al. Scale-recurrent network for deep image deblurring[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8174-8182.

[14] Yin Y, Jiang S, Robinson J P, et al. Dual-attention GAN for large-pose face frontalization[C]//2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020). IEEE, 2020: 249-256.

[15] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.

[16] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[17] Sagonas C, Panagakis Y, Zafeiriou S, et al. Robust statistical face frontalization[C]//Proceedings of the IEEE international conference on computer vision. 2015: 3871-3879.

[18] Liu S, Yang J, Huang C, et al. Multi-objective convolutional learning for face labeling[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3451-3459.

[19] Tian Y, Peng X, Zhao L, et al. CR-GAN: learning complete representations for multi-view generation[J]. arXiv preprint arXiv:1806.11191, 2018.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202401_10(1).0021

[20] Huang R, Zhang S, Li T, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2439-2448.

[21] Li P, Wu X, Hu Y, et al. M2FPA: A multi-yaw multi-pitch high-quality dataset and benchmark for facial pose analysis[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 10043-10051.