# Multi-Scale Gradients Self-Attention Residual Learning for Face Photo-Sketch Transformation

Shuchao Duan, Zhenxue Chen, Q. M. Jonathan Wu, *Senior Member, IEEE*, Lei Cai, and Dan Lu

***Abstract*— Face sketch synthesis, as a key technique for solving face sketch recognition, has made considerable progress in recent years. Due to the difference of modality between face photo and face sketch, traditional exemplar-based methods often lead to missed texture details and deformation while synthesizing sketches. And limited to the local receptive field, Convolutional Neural Networks-based methods cannot deal with the interdependence between features well, which makes the constraint of facial features insufficient; as such, it cannot retain some details in the synthetic image. Moreover, the deeper the network layer is, the more obvious the problems of gradient disappearance and explosion will be, which will lead to instability in the training process. Therefore, in this paper, we propose a multi-scale gradients self-attention residual learning framework for face photo-sketch transformation that embeds a self-attention mechanism in the residual block, making full use of the relationship between features to selectively enhance the characteristics of specific information through self-attention distribution. Simultaneously, residual learning can keep the characteristics of the original features from being destroyed. In addition, the problem of instability in GAN training is alleviated by allowing discriminator to become a function of multi-scale outputs of the generator in the training process. Based on cycle framework, the matching between the target domain image and the source domain image can be constrained while the mapping relationship between the two domains is established so that the tasks of face photo-to-sketch synthesis (FP2S) and face sketch-to-photo synthesis (FS2P) can be achieved simultaneously. Both Image Quality Assessment (IQA) and experiments related to face recognition show that our method can achieve state-of-the-art performance on the public benchmarks, whether using FP2S or FS2P.**

***Index Terms*— Face photo-sketch transformation, generative adversarial networks, residual learning, multi-scale gradients, self-attention mechanism, image-to-image translation.**

## I. Introduction

FORENSIC sketch face recognition is a critical technology in personal identification that was first proposed by Jain and Klare *et al.* [1]. Compared with photo-based face recognition, this area of study is more challenging and has great value in terms of practical application in law enforcement and criminal cases [2]. In law enforcement, the police usually search automatically in databases of criminals' face photos to find matches with photos of suspects, but photos of criminal suspects are often not available. Therefore, forensic sketches are crucial for tracking suspects. Under these conditions, forensic sketches drawn by artists based on witness descriptions or low-resolution photo-realistic captured by video surveillance are often used as substitutes in searching for suspects. However, it is difficult to match the photo-realistic with the corresponding sketch since they are in different modalities. A common solution to this issue is to perform face photo-to-sketch synthesis or face sketch-to-photo synthesis before matching and convert the two to the same modality, then implement the recognition task. Some work has been proposed for this effort. At present, most of the works focus on FP2S; a few research efforts involve FS2P, such as [3]. In this paper, the tasks of FP2S and FS2P are studied together.

The traditional exemplar-based methods [4]–[8] have made good progress after years of development. Their main processes include neighbor selection and reconstruction weight representation. They usually subdivide the test photos into patches with overlapping parts, match these test photo patches with the photo patches in the training set, and then use the corresponding sketch patches in the training set to synthesize the whole image. Although the exemplar-based methods have achieved good results in the synthesis of sketches, these synthesized images are too smooth to retain such fine content as forehead hair (see Fig. 1 Row 1). In addition, the process of patch matching and optimization is usually very time-consuming. Recently, some works using Convolutional Neural Networks (CNNs) have achieved great success in performing different image transformation tasks. Similar to [9] and [10], FS2P and FP2S are considered to be image transformation

Manuscript received July 11, 2020; revised September 10, 2020; accepted October 5, 2020. Date of publication October 15, 2020; date of current version November 6, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61876099, in part by the National Key Research and Development Program of China under Grant 2019YFB1311001, in part by the Scientific and Technological Development Project of Shandong Province under Grant 2019GSF111002, in part by the Shenzhen Science and Technology Research and Development Funds under Grant JCYJ20180305164401921, in part by the Foundation of Ministry of Education Key Laboratory of System Control and Information Processing under Grant Scip201801, in part by the Foundation of Key Laboratory of Intelligent Computing and Information Processing of the Ministry of Education under Grant 2018ICIP03, and in part by the Foundation of State Key Laboratory of Integrated Services Networks under Grant ISN20-06. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Andrew Beng Jin Teoh. *(Shuchao Duan and Zhenxue Chen are co-first authors.) (Corresponding author: Zhenxue Chen.)*

Shuchao Duan and Dan Lu are with the School of Control Science and Engineering, Shandong University, Jinan 250061, China (e-mail: dscvincent1995@gmail.com; ludan1994@foxmail.com).

Zhenxue Chen is with the School of Control Science and Engineering, Shandong University, Jinan 250061, China, and also with the Shenzhen Research Institute, Shandong University, Shandong University, Shenzhen 518057, China (e-mail: chenzhenxue@sdu.edu.cn).

Q. M. Jonathan Wu is with the Department of Electrical and Computer Engineering, University of Windsor, Windsor N9B 3P4, Canada (e-mail: jwu@uwindsor.ca).

Lei Cai is with the School of Information Engineering, Henan Institute of Science and Technology, Xinxiang 453003, China (e-mail: cailei2014@126.com).

Digital Object Identifier 10.1109/TIFS.2020.3031386

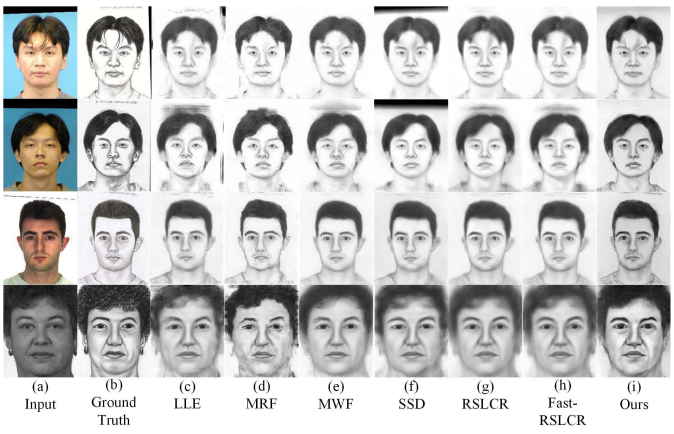



Fig. 1. Sketches synthesized by exemplar-based methods and our proposed method. The examples of the first three rows are selected from the CUFS database, and the last row is from the CUFSF database. From left to right: Input, Ground Truth, LLE, MRF, MWF, SSD, RSLCR, Fast-RSLCR, and ours. Sketches generated by exemplar-based methods are over smooth, which results in the blur effect on the contour edge. The results of the proposed method in (i) have more abundant facial texture.

tasks between photos and sketches. The emergence of the generative adversarial network has led to considerable progress in style transfer [11], image super-resolution [12], and image inpainting [13] tasks due to its powerful generative ability. Moreover, it has become a preferred method in modeling distribution and transformation between different domains. For example, in [9], a general framework using conditional GANs is proposed to solve the image transformation task in a supervised manner. Zhu *et al.* [10] define a non-linear mapping between the unpaired images based on the cycle consistency assumption. Although these methods can help to model the mapping relationship between the two domains, they only focus on a region of the size of the neighborhood convolutional kernel each time, which makes them have weak constraints on the overall content, thus neglecting the relationship between the features in the global perspective and leading to the inability to retain some of the detailed features in the synthesized image. These features are also important for the recognition task.

For our task, our ultimate goal is to improve the recognition accuracy, so it is very important to make full use of the dependencies between features in the global view. In particular, in the process of image reconstruction, the performance is mainly determined by the reconstruction quality of the facial region rather than the background region. As such, we should focus on strengthening the constraint of the features of the facial region. In order to preserve the relationship between the features in the process of image reconstruction, some works [14], [15] use decomposition structure to enlarge the size of the convolution kernel or introduce an effective encode layer at the top of the network to capture more abundant global context information. However, this reduces calculation efficiency and increases model complexity. As an effective way to model context, attention mechanism has achieved good results in many vision tasks. Class activation mapping (CAM) [16] is usually used to explore remote dependencies and is widely used in semantic segmentation tasks, but it always over activates the background region.

In this paper, our approach draws on the experience of the predecessors and is based on their ideologies, but there are some key differences. Based on the task of face photo-sketch transformation, we propose a multi-scale gradients self-attention residual learning framework that takes advantages of GAN and the self-attention mechanism.

In the process of model training, in order to increase the stability of the discriminator, we introduce a multi-scale gradients technique that allows to look at multi-scale outputs of the generator rather than a single output to improve the overlap between the real and false distributions to feedback more useful information to the generator. We design a deep residual network based on U-net architecture as our generator in order to retain more channel information and position information through skip connection. More importantly, the fusion of spatial self-attention and channel self-attention is used to make full use of the long-distance relationship between features and to selectively emphasize information features. In addition, the adversarial loss and perceptual loss based on the VGG-Net [17] are used to make the generated image more realistic. The main contributions of this work can be summarized as follows:

- We propose a multi-scale gradients self-attention residual learning framework for face photo-sketch transformation that can realize the mutual conversion between the source domain and the target domain.
- In the process of image transformation, the use of self-attention mechanism strengthens the constraint of facial features and is more robust to the interference of background and other factors, and further improves the quality of generated images.
- The combination of global residual connection and local residual connection ensures that the information of original features is not destroyed, and the loss of hair and other details due to the influence of self-attention distribution is avoided.
- Extensive experiments prove the effectiveness and superiority of the proposed method. This method's image quality evaluation and face recognition can achieve state-of-the-art performance on public benchmarks.

The rest of the paper is arranged as follows. Section II. reviews the existing FP2S methods and the development of image-to-image translation and the attention mechanism. Section III. introduces the proposed method. Section IV. gives the detailed experimental results and analysis. Section V. makes conclusions based on these.

## II. Related Works

This section will review the methods of face sketch synthesis, image-to-image translation, and the attention mechanism.

### A. Face Photo-Sketch Synthesis

The existing works on face sketch synthesis include exemplar-based methods and model-based methods. The exemplar-based method was originally put forth by Tang and Wang [18] using Principal Component Analysis (PCA) to construct sketches via linear transformation of the whole image, thus opening up the research of face sketch synthesis.

Subsequently, the texture and shape effect of the synthetic sketch are improved based on the PCA to further enhance the synthesized effect [19]. Liu *et al.* [4], inspired by Locally Linear Embedding (LLE), proposed to leverage k-nearest neighbor selection and least square loss function to get the reconstruction coefficient and estimate the nonlinear mapping from photo to sketch. To accelerate the image synthesis process, Song *et al.* [5] improved on the literature [4] to propose a spatial denoising method (SSD). Wang and Tang [6] proposed a method based on multi-scale Markov Random Field (MRF). Zhou *et al.* [7] improved MRF and proposed a synthesis method based on Markov Weight Field. Wang *et al.* [20] comprehensively surveyed the topic of face hallucination (FH) incorporating face image super-resolution and face sketch-photo synthesis and proposed several promising future directions and tasks, which helps readers to obtain a holistic understanding and deep insight into face sketch-photo synthesis. As can be seen, researchers have undertaken a lot of work in this area. Wang *et al.* [21] proposed a Bayesian face sketch synthesis method. Zhu *et al.* [8] proposed a DGFL framework, in which the deep features learned by depth network are weighted and integrated into a Bayesian network. Peng *et al.* [22] proposed a deep patch representation method for wild face photos based on a probabilistic graphical model. However, these methods are based on the assumption that the same photo patch will have the same corresponding sketch patch. However, due to the difference of modalities between them, the same photo patch does not always correspond to the same sketch patch, since the test sample is different from the training sample. This will lead to a loss of texture details in the synthesis of the sketch.

The other model-based methods benefit from the development of CNNs. Zhang *et al.* [23] proposed an end-to-end fully convolutional neural network structure to build the nonlinear mapping from photo to sketch. Although it can roughly estimate the shape of the face, however, the deep network can result in a blur effect due to using mean square error (MSE) loss. Based on Generative Adversarial Networks (GANs) [24], Wang *et al.* [25] proposed a post-processing framework via exemplar-based methods that can post-process the sketch generated by GAN to get a higher quality. And Lu *et al.* [26] proposed an FCN-based preprocessing framework in which using the exemplar-based methods can improve image quality. Based on style transfer [11], Chikontwe and Lee [27] proposed a method by combining the content loss and style loss to synthesize the face sketch. The methods of style transfer model their style as texture features, however, for face photo-sketch synthesis task, the sketch style with little texture is not suitable. Therefore, Yi *et al.* [28] proposed a hierarchical GAN model to transform a face photo into an APDrawing (Artist Portrait Drawing), which can generate different key areas of face separately to obtain more precise and better effect results. Chen *et al.* [29] proposed a method based on pyramid column features that first utilizes CNN to generate the content image containing contour and crucial facial features and then renders textures and shadows. To adress the problem of insufficient training data, Zhu *et al.* [30] first introduced knowledge transfer to learn knowledge using teacher network trained on large amount of data in related task and transfer them to student network. Based on the problem of the utilization of mutual interaction between two opposite mappings being lacking, Zhu *et al.* [31] proposed a collaborative framework using a collaborative loss to tackle it. This strategy successfully enhances the performance of the model for synthesizing images. Wang *et al.* [3] added a multi-level adversarial network on the basis of GAN to improve the quality of the synthesized image. Zhang *et al.* [32] used multi-domain adversarial learning to train GAN to establish the mapping relationship between photo domain and sketch domain.

However, all of these methods are limited to the local receptive fields of CNNs. As such, the long-distance feature relationship in the global view cannot be well-utilized, and some details cannot be well-preserved.

### *B. Attention Mechanism*

Recently, a large number of papers have proved that introducing an attention mechanism into a network can improve the feature expression ability of the network model. Bahdanau *et al.* [33] first proposed a attention model to extract the global dependence of input and apply it to machine translation. At the same time, attention modules are gradually applied to vision tasks. Wang *et al.* [34] proposed a non-local Neural Network (non-local NN) to capture long-range relationships that could be applied for image and video tasks. Based on the covariance matrix of non-local block, Du *et al.* [35] designed a new loss function based on PCA to better interact the features in the channel dimension. Woo *et al.* [36] added a spatial attention module on the basis of SE-Net [37] to average and maximize the pooling of features between channels. Fu *et al.* [38] proposed DA-Net based on the combination of CBAM and non-local NN, which uses the self-attention mechanism in image segmentation and can achieve more accurate segmentation through the long-distance context. Zhou *et al.* [16] proposed a method of applying a Class Activation Map (CAM) in global average pooling to CNNs to find the most discriminative part of the target. Zhang *et al.* [39] put forward the idea of applying a self-attention mechanism to GAN that could generate details with hints from all feature positions. Kim *et al.* [40] used CAM to distinguish source domain and target domain in an unsupervised image transformation task in order to focus on identifying changes of image region.

### *C. Image-to-Image Translation*

Since Goodfellow *et al.* [24] proposed Generative Adversarial Networks (GANs), they have achieved great success in different image generation tasks, including style transfer [11], image super-resolution [12], and image inpainting [13].

Models use input and output data to learn the parameter mapping between input and output. For example, Isola *et al.* [9] proposed a general framework using conditional GANs to solve the pixel-to-pixel image translation task in a supervised way. Wang *et al.* [41] introduced the pix2pixHD model to synthesize a high-resolution image from a semantic label map. However, these methods often require paired image

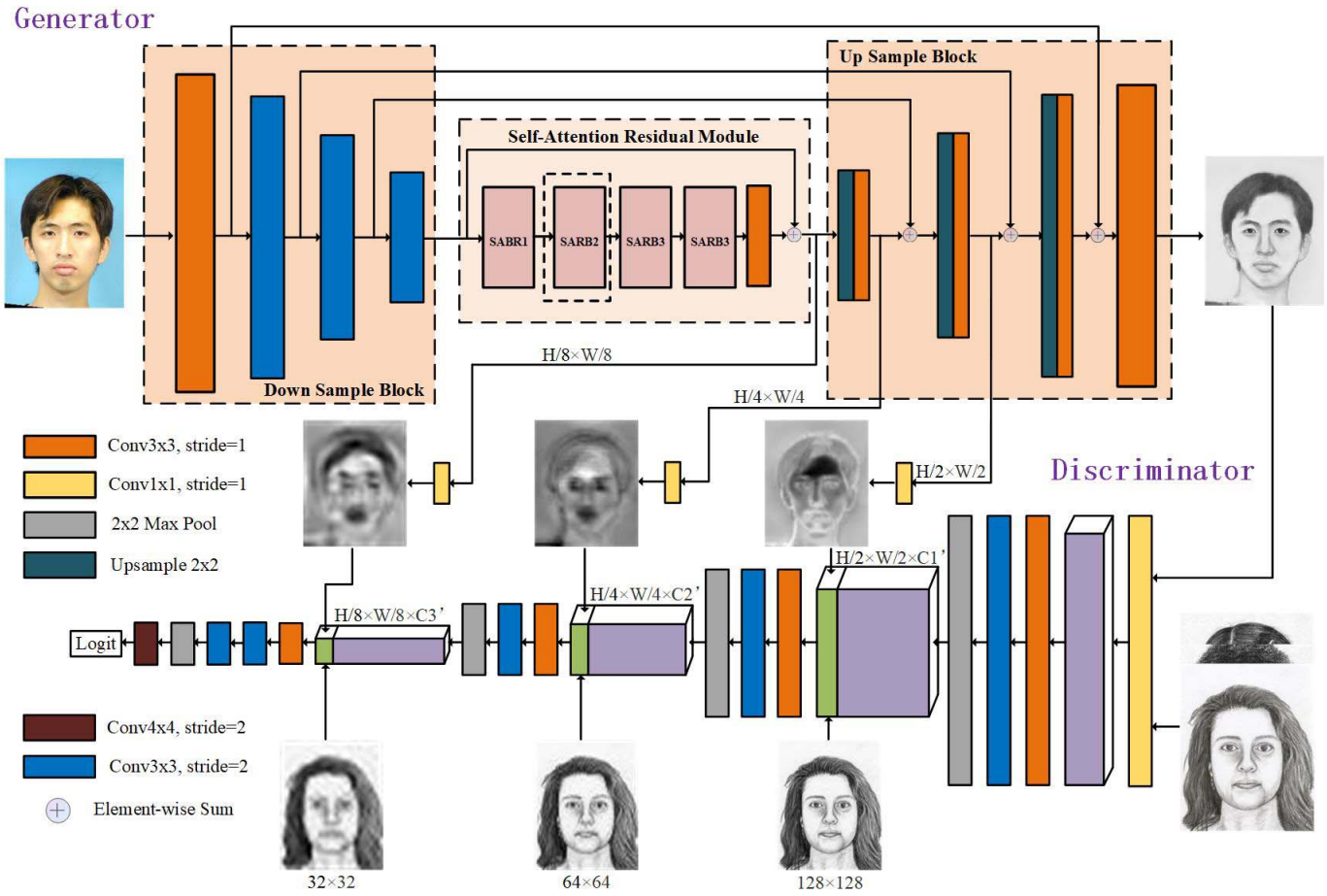



Fig. 2. The overall network architecture. Detailed description in Section III.A.

datasets in the training process. To overcome this limitation, Zhu *et al.* [10] defined a nonlinear mapping between unpaired images based on the cycle consistency hypothesis. Kim *et al.* [42] proposed a method of learning to find the relationship between domains and solved the problem of the lack of paired images in image-to-image translation settings through a model based on two different GANs coupling. Chen *et al.* [43] proposed a specific GANs-based method that can effectively learn to use the unpaired image set for training and establish the mapping of real-world photos and cartoon images. Because their methods discovered relationships between different domains, they can be used to successfully transfer styles.

## III. Proposed Methods

### A. Network Architecture

Our work aims to train a function $\mathbf{y} = G_{p\to s}(x)$, which maps images from the photo domain $X_p$ to the sketch domain $X_s$. Our framework consists of two generators ($G_{p\to s}$, $G_{s\to p}$) and two discriminators ($D_S$, $D_p$). The overall architecture of the proposed Multi-Scale Gradients Self-Attention Residual Network is illustrated in Fig. 2. Here, we only explain $G_{p\to s}$ and $D_s$ and vice versa.

Let $x \in X_p$ represent a sample from the photo domain. In our work, Generator $G_{p\to s}$ includes three parts: Down Sample Block $DS_p$, Up Sample Block $US_p$, and the Self-Attention Residual Module (SARM).

Because the deep convolutional layers will lead to the loss of some information and details of the features, we use U-Net as the backbone. The difference is that we apply the element-wise sum operation for the corresponding resolution features in the $DS_p$ and $US_p$ using skip connection to preserve more position information rather than concatenation. In the Self-Attention Residual Module, we stack self-attention residual blocks (detailed in Section III. B.) to adaptively weight the feature statistics and selectively emphasize information features.

Finally, in order to reduce the checkerboard effect, we use the combination of up sampling and convolutional operation instead of deconvolution to reconstruct the target image of different scales until it is equal to the input scale.

For the discriminator structure, on the basis of PatchGAN from [9], we modify it by adding the maximum pooling layer to reduce the redundant parameters and save memory. The detail of it is illustrated in Fig. 2.

### B. Multi-Scale Gradients Technique

In the training process of GAN, when the real distribution and the generated distribution are not overlapped enough, and because the discriminator has no ability to record historical

information, if discriminator only judge the final output of the generator, then the discriminator cannot feed back useful information to the generator.

To solve these problems, Cycle-GAN [10] uses an image buffer to store a specific number of previously generated images rather than the ones produced by the latest generators to increase the stability of discriminator and make sure that it cannot forget the information learned before. However, if the GPU is used for training, the data needs to be sent back to the CPU memory, and then transferred to the GPU memory after entering the image buffer, which leads to low computational efficiency. In our work, following the idea of MSG-GAN [44], the discriminator are allowed to ues not only the gradient of final output of the generator, but also the gradients of different resolutions outputs from the intermediate layer of the $\mathrm{US_p}$. The discriminator feeds back the judgement of multi-scale outputs to the generator to improve the overlap between the real and generated distributions, instead of just using a buffer to record historical information.

We define the function to generate outputs in different up sampling layers of the generator (see yellow block in Fig. 1) where the output corresponds to different down sample values of the final output image. We simply model $\gamma$ as $1 \times 1$ convolution, which transforms the intermediate feature image into an image.

$$O_i^g = \gamma_i \left( G_{p \to s}^i(x) \right) \tag{1}$$

where $i = 0, \ldots, 2$. $O_i^g \in R^{\frac{H}{2^{3-i}} \times \frac{W}{2^{3-i}} \times 1}$ is an image synthesized by the output of the $i^{th}$ layer of the $\mathrm{US_p}$ of the generator. $\gamma$, as a regularizer, requires that the learned feature mapping can be directly projected into the sketch domain.

Since the discriminator becomes a function of multiple scale outputs of the generator, the final discrimination loss function of the discriminator is not only the final output of the generator $G_{p \to s}$, but also the output of the upper sampling block middle layer $O_i^g$. We will use the real sketch domain image $\mathrm{y_s} \in \mathrm{X_s}$ (real sample) and generator final output image $y_g$ (fake sample) as input. $\delta_j$ is the middle layer function of the discriminator. The output of a corresponding to $j^{th}$ middle layer of discriminator $a_j$ is defined as:

$$a_{j+1} = \delta_j \left( \gamma' \left( \vartheta \left( O_i^g, a_j \right) \right) \right) \tag{2}$$

where $i + j = 4$ and $j = 2, 3, 4$. $\vartheta$ is a function that concatenates the output $O_i^g$ of the middle layer of the upper sampling block in the generator (or the down sampling version corresponding to the real sketch domain image $y_s$) and the corresponding output $\delta$ in the discriminator. $\delta_j$ is the function of channel dimension cascade, and $\gamma'$ is the same as $\gamma$, which is $1 \times 1$ convolution. The output of the final generated sample of discriminator is:

$$d_{fake} = D_s \left( y, \left[ O_0^g, \ldots, O_i^g \right] \right) \tag{3}$$

The output of the real sample of discriminator is:

$$d_{real} = D_s \left( y_s, \left[ y_o^s, \ldots, y_i^s \right] \right) \tag{4}$$

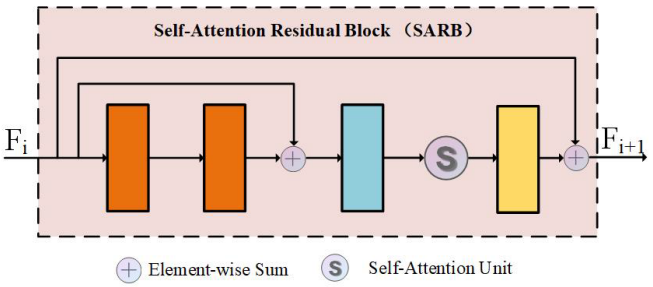



Fig. 3. Illustration of the self-attention residual module.

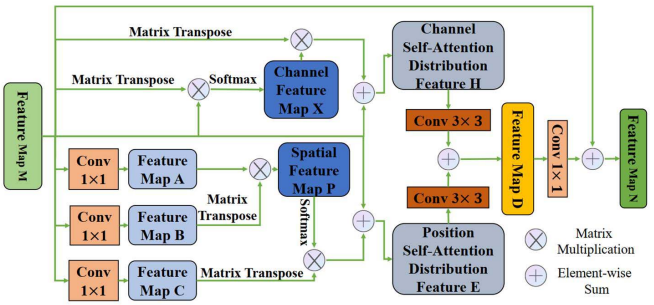



Fig. 4. Illustration of the self-attention unit.

### C. Self-Attention Residual Module

Extracting the correlation between the intermediate features is essential to enhance the discriminative learning ability of CNNs. Inspired by the previous work [45], we introduce a self-attention residual block (SARB) to explore and strengthen the constraints of adaptive reinforcement on facial features. SARB is mainly composed of two units, a residual learning unit (responsible for feature extraction) and a self-attention unit (responsible for attention distribution), as shown in Fig. 3.

In order to alleviate the problem of gradient vanish and explosion in deep network, we use global skip connection and local skip connection in the residual learning unit. The local residual block contains a basic residual block, and an additional $5 \times 5$ convolution operation and a self-attention unit are embedded in the global residual block. The residual learning unit can make full use of the low-frequency information and non-low-frequency information of the input image so as to greatly enhance the learning ability of the correlation between features. More importantly, the addition of global residual can keep the model from destroying the characteristics of the original features while retaining the important ones. In order to make our model learn to focus, we let the generator focus on the source domain image, which is more distinguishable from the target domain (such as glasses and mouth) while ignoring the background and other unimportant information; to do this, we added the self-attention unit, as shown in Fig. 4, SRAB can increase the proportion of important features adaptively via adding the self-attention unit.

Given $\mathrm{F} \in \mathbb{R}^{\mathrm{C} \times \mathrm{H} \times \mathrm{W}}$, we define the operation process of SARB with $\mathrm{SARB_i}(\cdot)$ function. In our work, $i = 1, 2, 3, 4$.

$$F_{i+1} = SARB_i \left( F_i \right) \tag{5}$$

In the self-attention unit, inspired by [38], we capture remote context on both spatial and channel dimensions. Among them, the position attention element encodes more extensive context information into local features so as to enhance their expression ability.

As shown in the lower part of Fig. 4, a local feature $\mathrm{M} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ obtained from the front layer is converted into two new feature maps A and B through the convolutional layer, where $\{\mathrm{A}, \mathrm{B}\} \in \mathbb{R}^{\mathrm{C}^*\times\mathrm{H}\times\mathrm{W}}$. Then, we reshape them to $\mathbb{R}^{C^*\times N}$, where $N = H \times W$ is the number of pixels and $C^* = \frac{c}{k}$. According to [39], we choose $\mathrm{k} = 8$. After that, we perform matrix multiplication between transpose of B and A to obtain the spatial feature map $\mathrm{P} \in \mathbb{R}^{\mathrm{N}\times\mathrm{N}}$.

$$p_{ji} = \frac{\exp\left(B_j^T \cdot A_i\right)}{\sum_{i=1}^{N} \exp\left(B_j^T \cdot A_i\right)} \tag{6}$$

Next, the softmax operation is applied to P to obtain $\mu_{ji}$:

$$\mu_{ji} = \frac{\exp\left(p_{ji}\right)}{\sum_{i=1}^{N} \exp\left(p_{ji}\right)} \tag{7}$$

This indicates the influence degree of position $i$ on position $j$. The more similar the feature representations of the two locations, the greater the correlation between them.

At the same time, we put M into the convolutional layer to generate a new feature map $\mathrm{C} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ and reshape it to $\mathbb{R}^{C\times N}$. Then, matrix multiplication is performed between the transpose of C and P, and the result is reshaped to $\mathbb{R}^{C\times H\times W}$. Finally, we multiply it by a scale parameter $\alpha$ and perform the sum of elements with feature M to obtain the final position self-attention distribution output $\mathrm{E} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$.

$$E_j = \alpha \sum_{i=1}^{N} \left(\mu_{ji}^T C_i\right) + M_j \tag{8}$$

where $\alpha$ is initialized to 0 and gradually learns to set more weights. It can be inferred from Eq.8 that the feature E of each position is the weighted sum of the features of all positions and the original features. Therefore, the feature information of the context can be selectively aggregated according to the spatial attention distribution in the global view.

In FP2S task, the input is RGB image, so the dependencies of each channel are very important for the quality of image synthesis. Each high-level channel map can be used as a response of a specific class, and different semantic responses are related to each other. By using the interdependencies between channel maps, we can highlight the feature map with scale dependency and improve the feature representation of specific semantics. Therefore, we use the channel attention, as shown in the top half of Fig. 4, to simulate the dependency between channels. Different from the spatial attention, we calculate the channel feature map $\mathrm{X} \in \mathbb{R}^{\mathrm{C}\times\mathrm{C}}$ directly from the original feature $\mathrm{M} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$. Specially, we reshape M to $\mathbb{R}^{C\times N}$ and then calculate matrix multiplication between the transpositions of M and M. Finally, we use the softmax layer to obtain the channel feature map $\mathrm{X} \in \mathbb{R}^{\mathrm{C}\times\mathrm{C}}$.

$$x_{ji} = \frac{\exp\left(M_j^T \cdot M_i\right)}{\sum_{i=1}^{C} \exp\left(M_j^T \cdot M_i\right)} \tag{9}$$

where, $x_{ji}$ indicates the influence of the $i^{th}$ channel on the $j^{th}$ channel. In addition, we multiply the X and M matrices and reshape their result to $\mathbb{R}^{C\times H\times W}$. Then, we multiply the result by a scale parameter $\beta$ and perform an element sum operation with feature M to obtain the final channel self-attention distribution output $\mathrm{H} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$:

$$H_j = \beta \sum_{i=1}^{N} \left(x_{ji} M_i\right) + M_j \tag{10}$$

where $\beta$ starts to learn weight gradually from 0. The final feature of each channel is the weighted sum of all channel features and original features, which helps to improve the features' distinguish ability. Finally, we apply $3\times3$ convolution operation $\varepsilon$ for H and E and then add the position attention distribution map and channel attention distribution map according to the elements to get the self-attention distribution feature D after $1 \times 1$ convolution operation $\gamma$. Then, we add them with M according to the elements to obtain the final feature map N.

$$N_j = \gamma\ \left(\varepsilon\left(H_j\right) + \varepsilon\left(E_j\right)\right) + M_j \tag{11}$$

### *D. Loss Function*

The full objective of the model includes four loss functions, i.e., $\mathcal{L}_{GAN}$, $\mathcal{L}_{perceptual}$, $\mathcal{L}_{tv}$, and $\mathcal{L}_{cycle}$. Here, we use the least square GAN [46] to stabilize the training target. In the following, $\mathbf{y} = G_{p\to s}(x)$.

*1) GAN Loss:* This is used to match the distribution of the photo domain image with the distribution of the sketch domain image in order to force the discriminator to correctly distinguish real or fake sketches. Following CycleGAN [10], the formula is expressed as:

$$\mathcal{L}_{GAN}^{p\to s} = \left(\mathbb{E}_{x\sim X_s}\left[\left(D_s(x)\right)^2\right] + \mathbb{E}_{x\sim X_p}\left[\left(1 - D_s\left(\mathbf{y}\right)\right)^2\right]\right) \tag{12}$$

where, $x \in X_p$, $D_s$ as discriminator for sketch domain. When D maximizes the loss while G minimizing it, GAN loss makes the synthesized sketch more closer to the sketch domain.

*2) Perceptual Loss:* Here, we use the pretrained vgg19 [17] model as the feature extractor and put the generated image into it to compare the feature map of pool1 and pool2 with that of the ground truth image. Because L1 loss and L2 loss will cause a blur effect in the synthesized image, our model uses perceptual loss to measure the similarity between high-level structures. In Eq. 13, H, W, and C represent the height, width, and channel number for the feature map, respectively. N is the number of feature maps generated by the vgg19 feature extractor.

$$\mathcal{L}_{\mathrm{perceptual}}^{p\to s} = \sum_{i=1}^{N} \frac{1}{HWC} \left\| \Phi^i(G_{p\to s}(x_p)) - \Phi^i\left(x_s\right) \right\|_2^2 \tag{13}$$

where $i$ represents the $i^{th}$ pooling layer in the VGG network. In our work, $i = 2, 5$ and $N = 2$.

*3) Cycle Consistency Loss:* Following CycleGAN [10], we use cycle consistency based on cycle framework to convert the generated image to the photo domain to supervise the image transformation process. For photo domain image $x \in X_p$ we expect: $x \rightarrow G_{p\rightarrow s}(x) \rightarrow G_{s\rightarrow p}(\mathbf{y}) \approx x$. The objective is expressed as follows:

$$\mathcal{L}_{cycle}^{p\rightarrow s} = \mathbb{E}_{x\sim X_p}\left[\left|x - G_{s\rightarrow p}(\mathbf{y})\right|_1\right] \tag{14}$$

*4) Total Variation Loss:* We use the total variation loss to further improve the synthesis quality and reduce artifacts.

$$\mathcal{L}_{tv}^{p\rightarrow s}(\mathbf{y}) = \sum_{i,j}\left(\left(\mathbf{y}_{i+1,j} - \mathbf{y}_{i,j}\right)^2 + \left(\mathbf{y}_{i,j+1} - \mathbf{y}_{i,j}\right)^2\right) \tag{15}$$

where $\mathbf{y}(i, j)$ represents the intensity value of the synthesized image at $(i, j)$.

*5) Full Objective:* Finally, we jointly train generators and discriminators to optimize the final objective:

$$\min_{G_{p\rightarrow s}, G_{s\rightarrow p}} \max_{D_p, D_s} \lambda_1\mathcal{L}_{GAN} + \lambda_2\mathcal{L}_{\text{perceptual}} + \lambda_3\mathcal{L}_{tv} + \lambda_4\mathcal{L}_{cycle} \tag{16}$$

where $\mathcal{L}_{GAN} = \mathcal{L}_{GAN}^{p\rightarrow s} + \mathcal{L}_{GAN}^{s\rightarrow p}$ and other loss functions are defined in the same way.

## IV. Experiments and Analysis

In this section, we will verify the effectiveness of our proposed method for FP2S and FS2P tasks through experimental comparative analysis.

It is not a simple task to evaluate the quality of a synthesized image. We focus on three aspects: qualitative, quantitative, and recognition accuracy. For the quantitative research, structural similarity index (SSIM) [47] and feature similarity index (FSIM) [48] are used to evaluate the structural similarity and feature similarity between images. Fréchet Inception Distance (FID) [49], which is used to calculate the distance between the real image and the fake image at the feature level, is also needed. For the qualitative aspect, we use visual perception. In addition, we conduct a user study to evaluate them. Finally, we test the recognition accuracy of different methods based on null-space linear discriminant analysis (NLDA) [50].

### A. Face Sketch Database

Our work utilizes the CUHK Face Sketch Database (CUFS) [18] and the CUHK Face Sketch FERET Database (CUFSF) [51] for training and testing.

CUFS includes 188 faces from the CUHK student database, 123 faces from the AR database, and 295 faces from the XM2VTS database for a total of 606 faces. For each face, the sketch drawn by the artist is based on the photo taken from the front with a neutral expression under normal lighting conditions. The CUFSF database includes 1,194 faces from the FERET database [52]. There are more challenging than the images in the CUFS database because of the change of illumination and the exaggeration of face contours by artists for each photo of each person.

### B. Implementation Details

Our model is trained from scratch, and the generator and discriminator are updated alternately in each iteration. Instance Normalization is applied to the generator and discriminator. We use Adam with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ to optimize the model. The number of training epochs is 400, and the learning rate is initially set to $10^{-3}$. After 100 epochs, the learning rate decreases with the coefficient of $10^{-1}$ until $10^{-5}$. The learning rate is from $10^{-3}$ to $10^{-5}$, decaying with a coefficient of $10^{-1}$. The batch size is set to 1. In the generator, we use the PReLU function with parameters instead of ReLU. In addition, we apply LeakyReLU to the discriminator. We empirically set the hyper-parameters of the full objective function as follows: $\lambda_1 = 1, \lambda_2 = 10, \lambda_3 = 10^{-4}, \lambda_4 = 10$. Our model is implemented on PyTorch and trained on a single NVIDIA RTX 2080Ti GPU.

### C. Ablation Studies

We evaluate our proposed method via ablation experiments. The studies show that the proposed method of Multi-Scale Gradients Self-Attention Residual Learning (MSG-SARL) is effective. More specifically, five different ablation experiments were performed and compared with the proposed MSG-SARL method, i.e., MSG-SARL w/o Cycle, MSG-SARL w/o Self-Attention, MSG-SARL w/o MSG, backbone, backbone w/ MSG, and backbone w/ Self-Attention. MSG stands for multi-scale gradients technology, Self-Attention represents the self-attention unit in RSAB, Cycle represents the cycle framework, and backbone represents the framework after removing the three function modules. MSG-SARL = backbone + MSG + Self-Attention + Cycle. Moreover, adding cycle framework can realize bidirectional transformation of the photo domain and the sketch domain at the same time. Through the above experiments, we verify the effects of the multi-scale gradients technique, the self-attention mechanism, and the cycle framework on the proposed method.

As shown in Fig. 5 (b-e), taking the FS2P task as an example, after visualizing the self-attention feature map in SARB, we can see that, by stacking SARB, the self-attention unit can gradually help the generator focus more accurately on the facial feature regions, thus strengthening the constraint on the facial features. In terms of visual perception, compared with (g) and (f) in Fig. 5, we can also see that the face texture becomes rich and clear after adding the self-attention unit, which verifies the favorable effect of the self-attention mechanism in the synthesis of face photos to strengthen the facial feature constraints. Without the use of the self-attention unit, the facial features in the synthesized image are not clear; in particular, the edges of the eyes and mouth are blurred. Fig. 6 (e) shows that, in the FP2S task, the synthesized sketch will produce noise without the self-attention unit. And from the IQA in Tables I, II, and III, it can be concluded that the introduction of the self-attention unit improves the quality of the synthesized image in both the FP2S and FS2P tasks.

As shown in the last three rows of Fig. 6 (b, c, f, h) and Row 3 in Fig. 7, after adding MSG, the contour of the face in the synthesized image becomes clear; in particular,

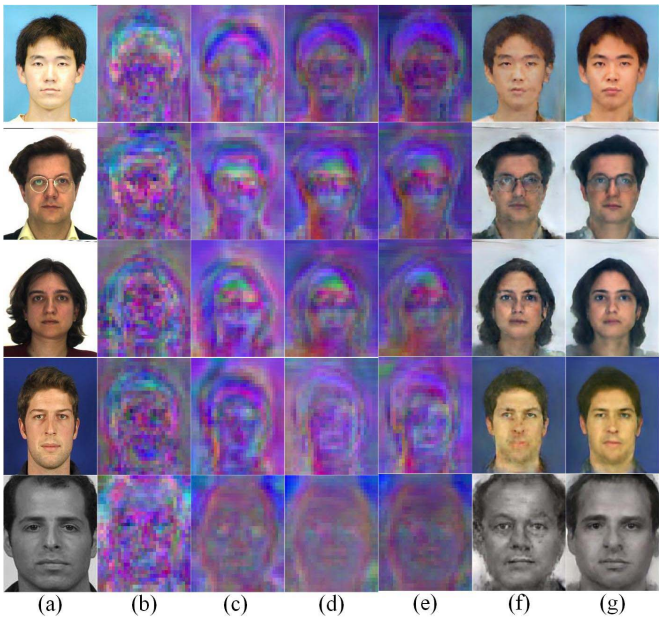



Fig. 5. Visualization of the self-attention feature maps in SARM and their effects shown in the ablation studies: (a) Ground truth; (b) self-attention feature map output by SARB1; (c) self-attention feature map output by SARB2; (d) self-attention feature map output by SARB3; (e) self-attention feature map output by SARB4; (f) our final results without the self-attention unit; (g) our final results with the self-attention unit. From top to bottom, the examples of the first four rows are selected from the CUFS database, and the last row is selected form the CUFSF database.

TABLE I
IQA VALUES (%) ON THE CUHK STUDENT DATABASE BASED ON DIFFERENT CONFIGURATIONS

| Configurations | | SSIM | FSIM |
|---|---|---|---|
| FP2S | backbone | 58.84 | 76.44 |
| | backbone w/ MSG | 59.45 | 76.63 |
| | backbone w/ Self-Attention | 60.60 | 76.90 |
| | MSG-SARL w/o MSG | 61.27 | 76.73 |
| | MSG-SARL w/o Cycle | 59.94 | 76.52 |
| | MSG-SARL w/o Self-Attention | 59.66 | 76.64 |
| | **MSG-SARL** | **61.64** | **77.11** |
| FS2P | MSG-SARL w/o MSG | 65.20 | 78.54 |
| | MSG-SARL w/o Self-Attention | 62.74 | 77.63 |
| | **MSG-SARL** | **66.07** | **78.88** |

the artifacts on the border of the chin region are eliminated. As shown in Tables I, II, and III, SSIM and FSIM values can be improved in CUHK, CUFS, and CUFSF after adding MSG, which indicates that MSG plays a positive role in improving the quality of composite images.

In addition, from Tables I, II, and III, it can be concluded that the introduction of the cycle framework also improves the IQA to a certain extent, and the constraint reconstruction image more closely matches the source domain image. Moreover, based on the cycle consistency, the mapping relationship between the two domains can be established at the same time. Thus, the FS2P task can also be achieved, as shown in Fig. 7.

### *D. Comparison With State-of-the-Art*

In order to prove the validity of our model, we use the other six methods, FCN [23], PS$^2$MAN [3], pix2pix [9], CycleGAN [10], DRIT [53], and U-GAT-IT [40], to make a qualitative and quantitative comparison with our proposed method in FP2S. Since there is less work like PS$^2$MAN that can also realize the FS2P task, in order to further illustrate the effectiveness of our method in photo-sketch transformation, we also compare our results with CycleGAN, DRIT, and U-GAT-IT, which can realize the transformation between the two domains. As a result of the above method, we use their open source code to train under the same conditions and test it after 400 epochs. The size of all training samples is resized to $256 \times 256$ for comparison with the ground truth and all test results are saved at the size of $200 \times 250$.


Fig. 6. Synthesized sketches of ablation studies. (a) Sketch domain ground truth; (b) backbone; (c) backbone w/ MSG; (d) backbone w/ self-attention; (e) MSG-SARL w/o Self-Attention; (f) MSG-SARL w/o MSG; (g)MSG-SARL w/o Cycle; (h) MSG-SARL(our proposed method). From top to bottom, the examples of the first four rows are selected from the CUFS database, and the last three rows are selected form the CUFSF database.

TABLE II
IQA VALUES (%) ON THE CUFS DATABASE BASED ON DIFFERENT CONFIGURATIONS

| Configurations | | SSIM | FSIM |
|---|---|---|---|
| FP2S | backbone | 49.53 | 75.58 |
| | backbone w/ MSG | 50.29 | 75.74 |
| | backbone w/ Self-Attention | 51.80 | 75.36 |
| | MSG-SARL w/o MSG | 52.51 | 75.63 |
| | MSG-SARL w/o Cycle | 50.75 | 75.62 |
| | MSG-SARL w/o Self-Attention | 50.48 | 75.80 |
| | **MSG-SARL** | **52.88** | **75.94** |
| FS2P | MSG-SARL w/o MSG | 61.56 | 78.36 |
| | MSG-SARL w/o Self-Attention | 59.00 | 77.79 |
| | **MSG-SARL** | **62.42** | **78.66** |

TABLE III

IQA VALUES (%) ON THE CUFSF DATABASE BASED ON DIFFERENT CONFIGURATIONS

| Configurations | | SSIM | FSIM |
|---|---|---|---|
| FP2S | backbone | 38.22 | 73.59 |
| | backbone w/ MSG | 38.71 | 73.76 |
| | backbone w/ Self-Attention | 41.31 | 73.66 |
| | MSG-SARL w/o MSG | 42.13 | 73.01 |
| | MSG-SARL w/o Cycle | 38.97 | 73.33 |
| | MSG-SARL w/o Self-Attention | 38.99 | **73.89** |
| | **MSG-SARL** | **42.30** | 73.16 |
| FS2P | MSG-SARL w/o MSG | 60.38 | 76.97 |
| | MSG-SARL w/o Self-Attention | 57.48 | 76.07 |
| | **MSG-SARL** | **61.14** | **77.34** |

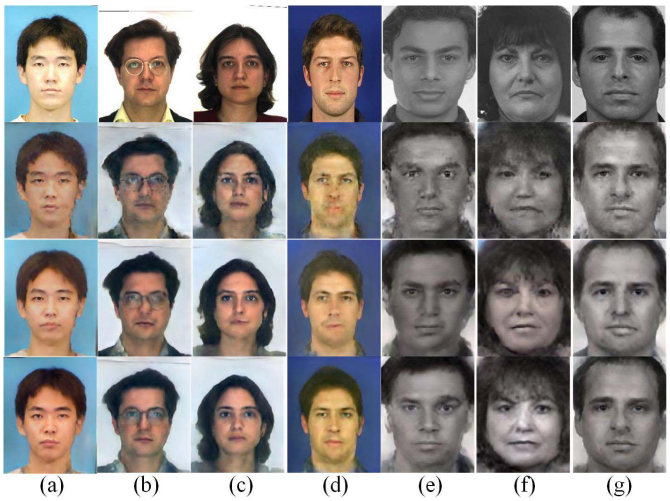



Fig. 7. Synthesized photos of ablation studies. Row 1: photo domain ground truth; Row 2: MSG-SARL w/o self-attention; Row 3: MSG-SARL w/o MSG; Row 4: MSG-SARL. (a-d) are selected from CUFS, and (e-g) are selected from CUFSF.

*1) Face Photo-to-Sketch Synthesis:* It can be observed from Fig. 8 that the FCN-based method has insufficient network feature extraction ability due to the lack of a deep network and that the use of MSE loss function leads to blurred output, while the methods based on GAN, pix2pix, and CycleGAN can eliminate the blurred effect by using the adversarial loss. But due to the instability in the training process, they tend to produce unexpected artifacts, such as the background of the first three rows in Fig. 8 (e) and the area around the mouth in the first two rows in Fig. 8 (f). PS$^2$MAN can improve the image quality after multi-level adversarial training, and the results of the second row in Fig. 8 (d) are better than those in Fig. 8 (f). However, in other databases, when there is a large difference between samples (such as background), the model will show a similar phenomenon to overfitting. After training on all databases, the model performs well on the CUHK database, but it does not perform well on other databases. As can be seen in the last five rows of Fig. 8 (d), the synthesized image becomes blurred compared with that in Fig. 8 (f). We call this phenomenon the deficiency of feature learning ability.

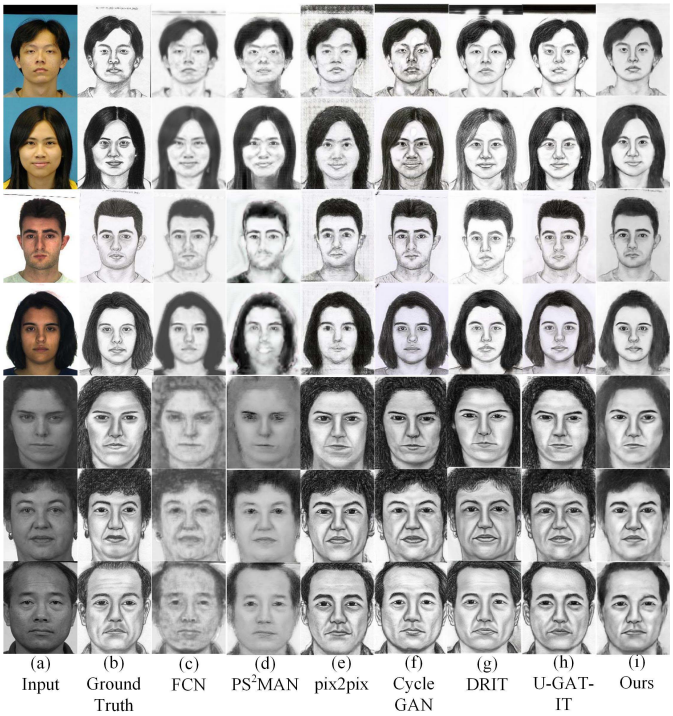



Fig. 8. Examples of face sketches synthesized by model-based methods and our proposed method. The first four rows are from CUFS, and the last three rows are from CUFSF. From left to right: Input, ground truth, FCN, PS$^2$MAN, pix2pix, Cycle-GAN, DRIT, U-GAT-IT, and ours.

TABLE IV

IQA VALUES (%) OF DIFFERENT METHODS FOR SYNTHESIZING FACE SKETCHES ON THE CUFS DATABASE

| Comparison Methods | SSIM | FSIM | FID |
|---|---|---|---|
| FCN | 52.14 | 69.36 | 80.97 |
| PS$^2$MAN | 47.83 | 70.91 | 94.43 |
| pix2pix | 30.84 | 66.29 | 153.81 |
| CycleGAN | 47.60 | 74.11 | 51.17 |
| DRIT | 45.68 | 71.38 | 54.02 |
| U-GAT-IT | 50.55 | 75.80 | **34.93** |
| **MSG-SARL(ours)** | **52.88** | **75.94** | t46.39 |

TABLE V

IQA VALUES (%) OF DIFFERENT METHODS FOR SYNTHESIZING FACE SKETCHES ON THE CUFSF DATABASE

| Comparison Methods | SSIM | FSIM | FID |
|---|---|---|---|
| FCN | 36.22 | 66.24 | 114.03 |
| PS$^2$MAN | 38.19 | 67.48 | 120.43 |
| pix2pix | 35.66 | 72.13 | 24.85 |
| CycleGAN | 34.24 | 71.13 | **17.81** |
| DRIT | 35.02 | 70.50 | 32.1 |
| U-GAT-IT | 36.46 | 72.54 | 21.47 |
| **MSG-SARL(ours)** | **42.30** | **73.16** | 38.25 |

In terms of visual perception, our method and DRIT, U-GAT-IT, and other image-to-image translation-based methods can achieve good results. However, as can be seen from Tables IV and V, our method can achieve the state-of-the-


Fig. 9. Examples of face photos synthesized by the methods of image-to-image translation and our proposed method. From top to bottom: photo domain ground truth, Cycle-GAN, PS$^2$MAN, DRIT, U-GAT-IT, and ours. (a-d) are selected from CUFS, and (e-g) are selected from CUFSF.

art performance in terms of SSIM and FSIM on both CUFS and CUFSF.

*2) Face Sketch-to-Photo Synthesis:* In the FS2P task, it can be seen from Row 2 in Fig. 9 that the color distortion of CycleGAN in the synthesis of photos does not completely convert the sketch domain image to the photo domain, but instead is based on the intermediate state between the two domains. One potential reason for this is that the L1 loss function is not used to recover the low-frequency information of the image in the process of training the network. However, due to the L1 loss and the L2 loss, the reconstructed image will be blurred. In the training process, our method uses the pretrained vgg19 as the feature extractor to extract the features of the low-level pooling layer in order to strengthen the constraint on low-frequency information to replace the L1 loss function and extract high-level pooling layer features to constrain the high-frequency information of the reconstructed image. PS$^2$MAN also encountered the same problem in the FP2S task when synthesizing photos. For example, in Fig. 9 Row 3 (a), the quality of synthesized images from the CUHK student database is significantly better than that from other databases.

Although DRIT and U-GAT-IT based on the image-to-image translation task have achieved good results in the FP2S task, there is also a phenomenon of insufficient feature ability learning when converting from sketch domain to photo domain. In the fourth row of Fig. 9, the results from the CUFS are seriously disturbed by skin color, background, and other factors, such as hair color and background color in (a), skin color in (b), and background color in (c). In the fifth row (a) of Fig. 9, there is also noise due to skin color interference. By contrast, our proposed method can not only retain rich facial texture details, but also produces relatively clear contour edges. More importantly, it can synthesize clear and high-quality face photos without the influence of sample differences between databases. As shown in Table VI, our method is second only to CycleGAN in terms of SSIM and FSIM on the CUFS database, while our method achieves the state-of-the-art performance on the CUFSF database, see Table VII.

TABLE VI

IQA VALUES (%) OF DIFFERENT METHODS FOR SYNTHESIZING FACE PHOTOS ON THE CUFS DATABASE

| Comparison Methods | SSIM | FSIM | FID |
|---|---|---|---|
| PS$^2$MAN | 42.59 | 64.51 | 182.51 |
| CycleGAN | **66.66** | **79.33** | 137.27 |
| U-GAT-IT | 56.64 | 75.955 | **54.65** |
| DRIT | 46.42 | 70.25 | 65.52 |
| **MSG-SARL(ours)** | 62.42 | 78.66 | 66.71 |

TABLE VII

IQA VALUES (%) OF DIFFERENT METHODS FOR SYNTHESIZING FACE PHOTOS ON THE CUFSF DATABASE

| Comparison Methods | SSIM | FSIM | FID |
|---|---|---|---|
| PS$^2$MAN | 52.06 | 75.13 | 83.50 |
| CycleGAN | 55.82 | 74.57 | **32.41** |
| U-GAT-IT | 56.13 | 75.21 | 33.84 |
| DRIT | 51.12 | 72.55 | 53.78 |
| **MSG-SARL(ours)** | **61.14** | **77.34** | 59.61 |

*3) User Study:* In order to further evaluate our method from the subjective judgement, we also conduct a user study to compare our results to U-GAT-IT [40], DRIT [53], and CycleGAN [10].

Here we learn from APDrawingGAN [28], we randomly selected 50 photo-sketch pairs from CUFS and CUFSF, including 20 pairs from CUFS and 30 pairs from CUFSF. There are four sketches synthesized by U-GAT-IT, DRIT, CycleGAN, and ours for each face photo. Then each image pair was expanded to a group of six images: one photo domain face image, one sketch domain ground truth, and four synthesized sketches. 30 volunteers were invited to take part in the study, each volunteer was randomly assigned to 10 groups from a total of 50 groups of images. After comparing with the photo domain face image and the sketch domain ground truth, the volunteer was asked to only choose the best one of four based on facial feature similarity and quality. The ranking results are summarized in Table VIII, in which U-GAT-IT is ranked the best in 36.33% of all cases, slightly higher than 28.67% of ours and 27.33% of CycleGAN. And our rank 1 ratio is very close to that of CycleGAN, which is about 20% higher than that of DRIT.

*4) Face Photo-Sketch Recognition:* Face photo-sketch recognition is an important application in the face photo-sketch

TABLE VIII

RANKING STATISTICS OF THE USER STUDY ON SYNTHESIZED SKETCHES. FOR EACH OF THE FOUR METHODS (U-GAT-IT, CYCLEGAN, DRIT, AND OURS), THE PERCENTAGE OF IT BEING RANKED BEST IS SUMMARIZED

| Methods | CycleGAN | U-GAT-IT | DRIT | Ours |
|---|---|---|---|---|
| Rank 1 | 27.33% | 36.33% | 7.67% | 28.67% |

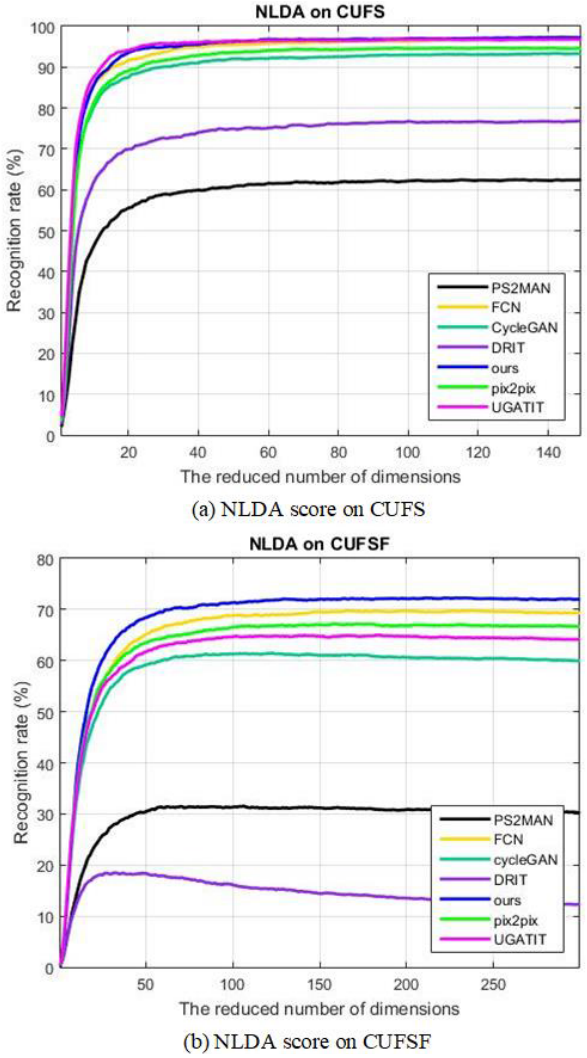



Fig. 10. Face sketch recognition accuracy against feature dimensions on CUFS and CUFSF for FP2S.

transformation task. We use NLDA to perform recognition experiments.

As can be seen in Fig. 10, in the recognition experiment on FP2S, our method achieves the highest recognition accuracy on both CUFS and CUFSF. In FS2P, our method is second only to CycleGAN on the CUFS database, while our method is superior to CycleGAN on the CUFSF database and achieves the highest recognition accuracy among all methods. This shows that our method has stable performance both on CUFS and CUFSF and that its generalization ability is better than those based on image-to-image translation task, such as CycleGAN.

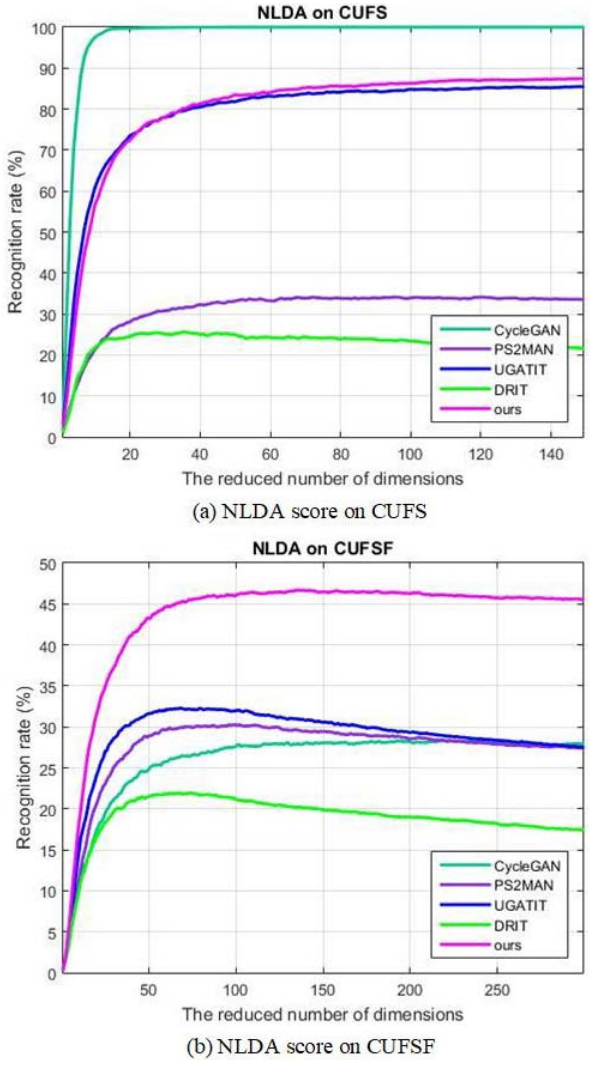



Fig. 11. Face photo recognition accuracy against feature dimensions on CUFS and CUFSF for FS2P.

## V. CONCLUSION

In this paper, we study face photo-sketch synthesis and face sketch-photo synthesis tasks and regard them as a face photo-sketch transformation task. A multi-scale gradients self-attention residual learning framework is proposed to achieve this task. The proposed method uses multi-scale gradients technology and deep residual network to solve the instability in the training process based on GAN. The self-attention unit is embedded in the residual block, which allows the generator to focus on the facial features and reduce the interference of the background area so that the image with richer texture information can be generated. At the same time, the perceptual loss based on vgg19 provides appropriate constraints on high-frequency information and low-frequency information in image reconstruction, which makes the generated image even more realistic. In addition, the forward and backward transformation processes are trained alternately by cycle framework based on cycle consistency loss. That is to say, in addition to minimizing adversarial loss and perceptual loss, the cycle consistency loss is also included in the objective

function. The evaluation experiments were conducted on the CUFS and CUFSF databases and compared with the recent state-of-the-art generation model. The experimental results show that our method can achieve great improvement in terms of the IQA and recognition accuracy of reconstructed images.

## REFERENCES


[1] A. K. Jain and B. Klare, "Matching forensic sketches and mug shots to apprehend criminals," *Computer*, vol. 44, no. 5, pp. 94–96, May 2011.

[2] B. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 639–646, Mar. 2011.

[3] L. Wang, V. Sindagi, and V. Patel, "High-quality facial photo-sketch synthesis using multi-adversarial networks," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 83–90.

[4] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 1005–1010.

[5] Y. Song, L. Bao, Q. Yang, and M. H. Yang, "Real-time exemplar-based face sketch synthesis," in *Computer Vision—ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 800–813.

[6] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.

[7] H. Zhou, Z. Kuang, and K. K. Wong, "Markov weight fields for face sketch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1091–1097.

[8] M. Zhu, N. Wang, X. Gao, and J. Li, "Deep graphical feature learning for face sketch synthesis," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, AAAI Press, Aug. 2017, pp. 3574–3580.

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

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[11] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015, *arXiv:1508.06576*. [Online]. Available: http://arxiv.org/abs/1508.06576

[12] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[13] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Jul. 2017, doi: 10.1145/3072959.3073659.

[14] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4353–4361.

[15] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.

[16] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[18] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 687–694.

[19] X. Tang and X. Wang, "Face sketch recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 50–57, Jan. 2004.

[20] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *Int. J. Comput. Vis.*, vol. 106, no. 1, pp. 9–30, Jan. 2014.

[21] N. Wang, X. Gao, L. Sun, and J. Li, "Bayesian face sketch synthesis," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1264–1274, Mar. 2017.

[22] C. Peng, N. Wang, J. Li, and X. Gao, "Face sketch synthesis in the wild via deep patch representation-based probabilistic graphical model," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 172–183, 2020.

[23] L. Zhang, L. Lin, X. Wu, S. Ding, and L. Zhang, "End-to-end photo-sketch generation via fully convolutional representation learning," in *Proc. 5th ACM Int. Conf. Multimedia Retr. (ICMR)*, New York, NY, USA: Association for Computing Machinery, 2015, pp. 627–634, doi: 10.1145/2671188.2749321.

[24] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2014, pp. 2672–2680.

[25] N. Wang, W. Zha, J. Li, and X. Gao, "Back projection: An effective postprocessing method for GAN-based face sketch synthesis," *Pattern Recognit. Lett.*, vol. 107, pp. 59–65, May 2018.

[26] D. Lu, Z. Chen, Q. M. J. Wu, and X. Zhang, "FCN based preprocessing for exemplar-based face sketch synthesis," *Neurocomputing*, vol. 365, pp. 113–124, Nov. 2019, doi: 10.1016/j.neucom.2019.07.008.

[27] P. Chikontwe and H. J. Lee, "Towards robust face sketch synthesis with style transfer algorithms," in *IT Convergence and Security 2017*, K. J. Kim, H. Kim, and N. Baek, Eds. Singapore: Springer, 2018, pp. 172–179.

[28] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "APDrawingGAN: Generating artistic portrait drawings from face photos with hierarchical GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10743–10752.

[29] C. Chen, X. Tan, and K.-Y.-K. Wong, "Face sketch synthesis with style transfer using pyramid column feature," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 485–493.

[30] M. Zhu, N. Wang, X. Gao, J. Li, and Z. Li, "Face photo-sketch synthesis via knowledge transfer," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1048–1054.

[31] M. Zhu, J. Li, N. Wang, and X. Gao, "A deep collaborative framework for face photo–sketch synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3096–3108, Oct. 2019.

[32] S. Zhang, R. Ji, J. Hu, X. Lu, and X. Li, "Face sketch synthesis by multidomain adversarial learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1419–1428, May 2019.

[33] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473

[34] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.

[35] Y. Du, C. Yuan, B. Li, L. Zhao, Y. Li, and W. Hu, "Interaction-aware spatio-temporal pyramid attention networks for action classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 373–389.

[36] S. Woo, J. Park, J. Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[38] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[39] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. 36th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 97, K. Chaudhuri and R. Salakhutdinov, Eds. Long Beach, CA, USA: PMLR, Jun. 2019, pp. 7354–7363.

[40] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," 2019, *arXiv:1907.10830*. [Online]. Available: http://arxiv.org/abs/1907.10830

[41] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[42] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," 2017, *arXiv:1703.05192*. [Online]. Available: https://arxiv.org/abs/1703.05192

[43] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "CartoonGAN: Generative adversarial networks for photo cartoonization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9465–9474.

[44] A. Karnewar and O. Wang, "MSG-GAN: Multi-scale gradients for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7799–7808.

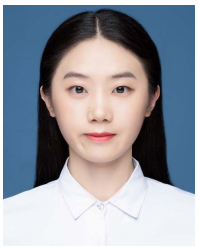
[45] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.
[46] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.
[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[48] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
[49] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," 2017, *arXiv:1706.08500*. [Online]. Available: https://arxiv.org/abs/1706.08500
[50] L.-F. Chen, H.-Y.-M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognit.*, vol. 33, no. 10, pp. 1713–1726, Oct. 2000.
[51] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. CVPR*, Jun. 2011, pp. 513–520.
[52] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, Apr. 1998.
[53] H. Y. Lee, H. Y. Tseng, J. B. Huang, M. Singh, and M. H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 35–51.

**Shuchao Duan** was born in Shandong, China, in 1995. He received the B.S. degree from the College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao, China, in 2018. He is currently pursuing the M.S. degree in control science and engineering with the School of Control Science and Engineering, Shandong University, Jinan, China. His current research interests include machine learning, deep learning, and sketch face synthesis.

**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. Since 2007, he has been an Associate Professor with the School of Control Science and Engineering, Shandong University. From 2012 to 2013, he was a Visiting Scholar with Michigan State University, East Lansing, MI, USA. 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 *Signal Processing: Image Communication* (SP-IC) journal. His research interests include image processing, pattern recognition, and computer vision, with applications to face recognition.

**Q. M. Jonathan Wu** (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Wales, Swansea, U.K., in 1990. He was with the National Research Council of Canada for ten years from 1995, where he became a Senior Research Officer and a Group Leader. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada. He has published more than 300 peer-reviewed articles in computer vision, image processing, intelligent systems, robotics, and integrated microsystems. His current research interests include machine learning, 3-D computer vision, video content analysis, interactive multimedia, sensor analysis and fusion, and visual sensor networks. He has served on technical program committees and international advisory committees for many prestigious conferences. He holds the Tier 1 Canada Research Chair of Automotive Sensors and Information Systems. He was an Associate Editor of IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS PUBLICATION INFORMATION and the *International Journal of Robotics and Automation*. He is currently an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and *Cognitive Computation* journal.

**Lei Cai** was born in Henan, China, in 1979. He received the B.S. degree from the Logistics Institute of Air Force, Xuzhou, China, in 2002, the M.S. degree from Xuzhou Air Force College, Xuzhou, China, in 2006, and the Ph.D. degree from Air Force Engineering University, Xi'an, China, in 2009. He has been an Associate Professor with the School of Artificial Intelligence, Henan Institute of Science and Technology. His research interests include image processing, pattern recognition, light field reconstruction, multiagent adaptive coordination.



**Dan Lu** was born in Shandong, China, in 1994. She received the B.S. and M.S. degrees from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2017 and 2020, respectively. Her current research interests include machine learning, deep learning, and sketch face recognition.