IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 17, 2020, accepted November 24, 2020, date of publication November 30, 2020,
date of current version December 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3041284

MHGAN: Multi-Hierarchies Generative Adversarial Network for High-Quality Face Sketch Synthesis

KANGNING DU, HUAQIANG ZHOU, LIN CAO, YANAN GUO, AND TAO WANG

Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, Beijing Information Science and Technology University, Beijing 100101, China
School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing 100101, China

Corresponding author: Lin Cao (charlin@bistu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671069, Grant 62001033, and Grant 62001034; in part by the Qin Xin Talents Cultivation Program of Beijing Information Science and Technology University under Grant QXTCP A201902; and in part by the General Foundation of Beijing Municipal Commission of Education under Grant KM202011232021.

ABSTRACT Face sketch synthesis has made significant progress in the past few years. Recently, GAN-based methods have shown promising results on image-to-image translation problems, especially photo-to-sketch synthesis. Because the facial sketch has a hyper-abstract style and continuous graphic elements, compared with other image styles, its local details are easier to expose small artifacts and blur. The existing face sketch synthesis methods lack models for specific facial regions and usually generate face sketches with coarse structures. To synthesis high-quality sketches and overcome the blurs and deformations, this paper proposes a novel Multi-Hierarchies GAN, which divides the face image into multiple hierarchical structures to learn different regions' features of the face. It includes three modules: a local region module, mask module, and fusion module. The local region module can learn the detailed features of different local regions of the face by GAN. The mask module can generate a coarse facial structure of a sketch and uses the facial feature extractor to enhance the high-level image and learn the latent spaces' feature. The fusion module can generate the final sketch by combining fine local regions and coarse facial structure. Extensive qualitative and quantitative experiments illustrate that the proposed method outperforms the state-of-the-art methods on the CUFS and CUFSF standard datasets and photos on the internet.

INDEX TERMS Face sketch synthesis, generative adversarial network, facial feature extractor, multi-hierarchies GAN.

I. INTRODUCTION

Face sketch synthesis is the process of generating face sketches from face photos. Face sketch synthesis has been studied for a long time due to its wide application. It plays an essential role in digital entertainment [1] and law enforcement based on video surveillance [2]. In law enforcement and criminal cases, the intelligent security system [3] can automatically retrieve photos of suspects from the police face database, so that the judicial authorities can quickly narrow down the scope of potential suspects. In practice, suspects' photos are usually hard to acquire, and police sought the commercial software or experienced artists to generate sketches of a suspect based on the description of an eyewitness. Other than the applications in security, face sketch synthesis also has several applications in digital entertainment. It has also become increasingly popular among smartphone users and social networks, where sketches are used as profile photos or avatars. Thus, face sketch synthesis is an important practical problem.

In the past decade, various methods have been proposed to achieve high-quality face sketch synthesis. These methods can be divided into two categories: data-driven methods and model-driven methods. The data-driven methods first perform image block processing on the training data, perform nearest neighbor selection and linear combination

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

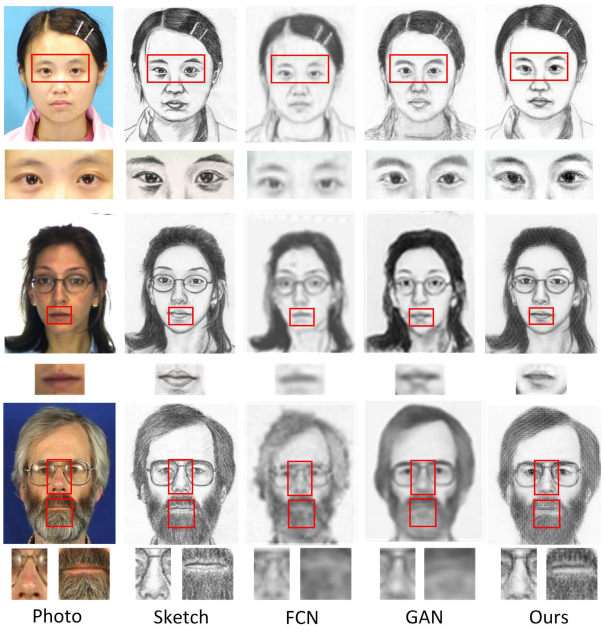

FIGURE 1. Comparison results of different methods on different sketch standard databases.

weight calculation, and finally select the best image block for sketch stitching. Since the synthesized sketch block is a linear combination of the training set sketch block, the data-driven methods can obtain good textures and facial detail features. Wang *et al.* [4] computed the linear combination coefficients and red used them to construct the target sketch by projecting the testing photos onto the training photos. However, the data-driven methods are time-consuming in the nearest neighbor selection process of the image block. The model-driven methods focus on learning the mapping relationship between photos and sketches from the training photo-sketch pairs offline. In the test phase, the model-driven methods can directly transform photos into sketches without searching in the training data, and the synthesis speed is faster than data-driven methods. Thanks to the breakthroughs in the model-driven methods, Generative Adversarial Networks (GAN) have received widespread attention in image style transfer. Zhang *et al.* [5] combined the high-frequency features of samples on the results of GAN to refine the texture.

Compared with other photo styles, sketches have more substantial semantic constraints. The loss and movement of facial features are subjectively unacceptable (even small local defects, such as around the eyes). Due to image elements of different facial regions (such as eyes and hair) in the sketch are inconsistent, it is difficult for a single network to learn multiple regions' features. The state-of-the-art methods based on model-driven can generate barely satisfactory results. However, these methods have not specified the targeted network for different facial regions. Therefore, these methods do not capture the facial detail well. Noise and deformation still exist in the synthesized results. Figure.1 shows the comparison of the synthesis results of the common method and the proposed method. The red box in Figure.1 shows that the FCN and GAN methods cannot generate subtle facial textures, and blur areas can be found in the local regions. It can illustrate the limitation of face sketch synthesis about these methods.

To address the above challenges, we propose a Multi-Hierarchies GAN (MHGAN) for face sketch synthesis. The proposed MHGAN is divided into local region module, mask module, and fusion module. The local region module divides the input photo into multiple hierarchies, and each hierarchy uses a GAN model to capture facial features and generate a corresponding sketch. Specifically, each hierarchy network is designed with an independent loss function to reduce noise and detail loss in the synthesis process. The model architecture in the proposed method can enhance the sketches' shadow and light and draw delicate contour curves. The mask module uses the facial feature extractor to calculate the *mask_feature* loss between the synthesized sketch and the real sketch called MF loss. We also use the smooth representation of the input image, which are photo's high-level features extracted by a white box function. Then we calculate the *mask_structure* loss of the synthesized sketch and the smooth image called MS loss. The mask module uses adversarial learning to provide mask sketch for face sketch synthesis and adds a controlling factor to ensure sufficient training of the generation network and the discriminant network. The fusion module uses the local region module's results, mask module's result and landmarks to generate the final sketch. To prove the effectiveness of the proposed method, we have performed extensive experiments on the Chinese University of Hong Kong face sketch benchmark dataset (CUFS) [6] and the CUHK face sketch face recognition technology dataset (CUFSF) [7] and photos on the internet. Compared with state-of-the-art methods, the experimental results show that the proposed method has more key facial details.

The main contributions of this paper are as follows:

1) We propose a novel end-to-end MHGAN framework for face sketch synthesis, which can generate high-quality and expressive clear face sketch images. In particular, our method can be applied to a variety of styles and different races.

2) The artist uses a variety of graphic elements for the face regions when creating a sketch. In order to synthesize a better sketch, the proposed framework divides the face into multiple hierarchies, each of which is controlled by separated loss functions. We also use a facial feature extractor to extract high-level features and texture details. The network introduces a total loss function, making the model more suitable for face sketch synthesis tasks.

3) We conducted multiple sets of comparative experiments in the experimental section. The ablation experiment compares the qualitative and quantitative results of different components of the proposed method to illustrate the contribution of each component; The comparison with traditional face sketching methods synthesis show that the proposed method can generate sketch images with finer textures; The comparison with GAN-based methods show that the contribution

made by the proposed method can be more suitable for the face sketch synthesis; The comparison with the state-of-the-art face sketch synthesis methods show that the proposed method has better performance in face sketch synthesis. The above comparative experiments have been carried out qualitative and quantitative comparative experiments in multiple sketch face datasets and real world's photos, which have proved the proposed method's outperforms these methods in addressing the problems of blurring facial features and losing facial details.

The rest of this paper is organized as follows. Section II introduces the current representative to face sketch synthesis methods. Section III introduces the proposed method in detail. Section IV presents various experimental results and comprehensive analysis. Section V gives a summary of this paper.

II. RELATED WORK

In this section, we discuss some existing face sketch synthesis methods. These methods could be roughly divided into two main types: the data-driven methods and the model-driven methods. In addition, the GAN-based synthesis methods in model-driven will be introduced.

A. DATA-DRIVEN FACE SKETCH SYNTHESIS METHODS

The data-driven methods usually use training set patches for linear combination to generate red face sketches. Tang and Wang [8] first proposed a face sketch synthesis method using feature transformation, but this method will cause local details to be lost or textures to be too smooth. Liu *et al.* [9] used the Local Linear Embedding (LLE) method to synthesize the face sketch. With the introduction of the Markov model, it has been widely used in many fields. For example, such as Peng *et al.* [10]–[12] proposed various applications of Markov-based jumping systems (MJSs). In face sketch synthesis, Wang and Tang [6] proposed to use a Markov Random Field (MRF) model to construct a data compatibility function and a smooth compatibility function, and select the most similar image block of target sketch for synthesis. Zhou *et al.* [13] used Markov Weighted Field (MWF) for face sketch synthesis, thereby improving the synthesized result. Gao *et al.* [14] used the sparse representation to measure the linear combination weights of similar training patches and then decompose the face into a sparse coefficient matrix and dictionary. Zhang *et al.* [15] used the sparse coefficient matrix to search candidate image blocks for test photos. Because the use of a sparse coefficient matrix can effectively reduce the computational complexity, but the lack of local constraints will lose facial information. Wang *et al.* [16] proposed a comprehensive face sketch synthesis method through simple offline Random Sampling and Local Constraints (RSLCR). The main advantage of the above method is that it can synthesize facial details from the linear combination of training image blocks. However, the large amount of calculation of the above method leads to poor real-time performance and low practicability.

B. MODEL-DRIVEN FACE SKETCH SYNTHESIS METHODS

The model-driven methods aim to learn the mapping relationship between face photos and face sketches offline. Chang *et al.* [17] proposed a face sketch synthesis method using ridge regression and correlation vector machine to learn the mapping relationship between photo-sketch patch pairs. Zhu *et al.* [18] divide the training photo-sketch patch pairs into clusters. These clusters can learn the mapping relationship by using a simple ridge regression model. Zhang *et al.* [19] used Fully Convolutional Network (FCN) to synthesize face sketches. Although the detailed information of a specific identity can be preserved, since only convolutional layers are stacked in the network, the real texture details will still be lost. Sheng [20] proposed a deep neural representation guidance method using enhanced 3D patch matching and cross-layer cost aggregation. These methods are trained offline, and the sketch results can be quickly obtained in the test phase. However, the model-driven method's final results often lack facial details or blurry artifacts.

C. GAN BASED FACE SKETCH SYNTHESIS METHODS

GAN has developed rapidly and has been widely used in the image style transfer field. Isola *et al.* [21] proposed a conditional GAN to learn the mapping relationship between input images and output images. This mapping can be applied to various style transfer tasks (e.g., labels to the street scene, aerial to map, day to night, and edges to photo). Zhang *et al.* [22] added a probabilistic graphic model to the GAN-based structure and proposed a method for synthesizing face sketches from coarse to fine. Besides, Zhang *et al.* [1] imitated the painter's painting process and added more detailed parts to the work of [22] to achieve better results. To solve the problem that the unpaired training set cannot be used, Zhu *et al.* [23] proposed using a cycle-consistent GAN (Cycle-GAN) to learn the image transfer mapping of unpaired images. Similar idea can be found in the latest work of PS^2-MAN [24]. Zhang *et al.* [25] proposed a new face sketch synthesis method by using Multi-Domain Adversarial Learning (MDAL). Chen *et al.* [26] proposed an example-based method (FSW) to subdivide the feature map of the input photo into overlapping small pieces, and then use the corresponding small sketch pieces in the feature space to form a fake sketch feature representation.

Although the face sketch synthesis technology has achieved remarkable results, the existing methods do not consider the end-to-end model problem of region facial details. There are still smooth, blur contours and small artifacts in the facial local region's synthesized sketch. To solve the above problems, we combined the GAN (realistic visual effects) and multi-hierarchies end-to-end framework (used to solve rough facial areas) to study the face sketch synthesis method.

III. PROPOSED METHOD

A. NOTATION

As shown in Figure 2, the proposed MHGAN framework is divided into the local region module, mask module, and fusion module. MHGAN models the process of learning to transform face photos domain P to the face sketch domain S as a function Ψ. The process can be expressed as: $S' = \Psi(P)$. MHGAN divides the face photos into five hierarchical structures $local\_parts = \{eye\_l, eye\_r, nose, mouth, mask\}$ as lp. In the local region and mask module, networks learn from the paired training set $Z_{Data} = \{(p_i, s_i)|p_i \in P, s_i \in S, i = 1, 2 \ldots, N\}$, where N is the number of photo-sketch pairs in the training set. In the fusion module, MHGAN uses the synthesized results of local region and mask module and landmarks to generate fake sketch. The fusion process can be expressed as function T: $s'_i = T(s'_{lp,i}, landmark)$.

The proposed framework consists of multiple generators and discriminators, all of which are CNNs networks, designed explicitly for multi-hierarchies face structures. The hierarchical generator of the proposed network is defined as G. The generator G contains multiple partial generators:$G = \{G_{eye\_l}, G_{eye\_r}, G_{nose}, G_{mouth}, G_{mask}\}$. The hierarchical discriminator is generally defined as the discriminator D, and the five hierarchical structures images of the face are input into different discriminators: $D = \{D_{eye\_l}, D_{eye\_r}, D_{nose}, D_{mouth}, D_{mask}\}$. Different local discriminators will discriminate different local sketches to evaluate local regions characteristics.

MHGAN divides the face photos into multiple hierarchical structures because the artist uses different painting techniques for different regions of the face during the drawing process. A face sketch usually spends more time on the eyes, such as pupils, eye corners, etc. When drawing the eyes, the artist will use short and powerful brush strokes. The image elements drawn for the mouth will usually follow the upper and lower lip lines and fill the shape with thin lines. Therefore, the facial region features' drawing process will be quite different. The standard GAN uses a single generator to synthesize the entire face sketch, and all facial regions share generator parameters, making it difficult to generate all facial region features properly. Therefore, MHGAN's hierarchical network design with multiple GANs can help the model better learn facial features in different positions and generate high-quality face sketches.

MHGAN can get fine local sketches and coarse mask sketches by the local region and mask module, and MHGAN inputs them into the fusion module to generate fake sketches. There may be inconsistent boundaries in the fusion process, making it subjectively difficult to accept the semantic information of the synthesized sketch. The fusion module proposes to use the non-conservative guidance field of the foreground sketch and the background sketch to solve the boundary inconsistency problem in the fusion process.

B. LOCAL REGION MODULE

In order to better learn the facial features in different regions of the input image, the local region module includes four local regions $lp \in \{eye\_l, eye\_r, nose, mouth\}$. Multiple generators can extract sketch features of local areas that preserve the artist's drawing style. Put the local generator G_{lp} and the local discriminator D_{lp} into four local networks respectively. After the model is trained, each local generator can transform the facial region photo $p_{lp,\,i}$ into the corresponding facial region sketch $s'_{lp,i}$. The local generator network is constructed by the modified U-Net. Each of G_{eye_l}, G_{eye_r}, G_{nose}, and G_{mouth} is a U-Net with three down-convolution and three up-convolution blocks. A U-Net with skip connections can incorporate multi-scale features, such as low-level features, and provide sufficient but not excessive flexibility to learn the artist's drawing techniques for different facial regions in sketches. Its input is as follow:

$$s'_{lp\,,\,i} = G_{lp}(p_{lp\,,\,i}), \tag{1}$$

where $p_{lp,\,i}$ is the input of the hierarchical generator G_{eye_l}, G_{eye_r}, G_{nose}, and G_{mouth}.They are local regions centered on the facial landmarks (i.e., left eye, right eye, nose, and mouth) obtained by the MTCNN [27], and the region images size is $h \times w$.

In the local region module, the loss function L_{local_adv} can help the discriminator better and correctly distinguish the authenticity of the input image. This module uses the cross-entropy loss in the Cycle-GAN method as the L_{local_adv} for adversarial loss and is defined as:

$$L_{local\_adv} = \sum_{D_{lp} \in D} \begin{matrix} E_{(p_{lp\,,\,i},\, s_{lp\,,\,i}) \sim Z_{data}} [\log D_{lp}(p_{lp,i}, s_{lp,i}) \\ + \log(1 - D_{lp}(p_{lp,i}, G_{lp}(p_{lp,i})))] \end{matrix}. \tag{2}$$

For each $D_{lp} \in \{D_{eye\_l}, D_{eye\_r}, D_{nose}, D_{mouth}\}$, the input images $p_{lp,i}$, $s_{lp,i}$ and $G_{lp}(p_{lp,i})$ are all matches to the local region specified by D_{lp}. When the discriminator D_{lp} maximizes L_{local_adv} and G_{lp} minimizes this loss, L_{local_adv} will make the synthesized sketch closer to the target domain S.

In the local region module, a strict L_1 loss is set in each hierarchical structure. The four local regions photos are $p_{eye\_l,i}$, $p_{eye\_r,\,i}$, $p_{nose,\,i}$, and $p_{mouth,\,i}$ respectively, and their loss function is defined as:

$$\begin{aligned} L_{local\_l_1} = E_{(p_{lp,i}, s_{lp,i}) \sim Z_{data}} [& \left\| G_{eye\_l}(p_{eye\_l,i}) - s_i) \right\|_1 \\ & + \left\| G_{eye\_r}(p_{eye\_r,i}) - s_{eye\_r,i} \right\|_1 \\ & + \left\| G_{nose}(p_{nose}) - s_{nose,i} \right\|_1 \\ & + \left\| G_{mouth}(p_{mouth,i}) - s_{mouth,i} \right\|_1 . \end{aligned} \tag{3}$$

The local discriminator D focuses on distinguishing whether the generated "fakes" sketches is the real sketch. Except for the difference in the input image size of different discriminators, the network structure is the same as Patch-GAN [21]. The Patch-GAN discriminator takes a 70×70 patch as the input image to avoid the lack of high-frequency information caused by directly inputting the two complete images and examines the style of each patch. Different facial partial patches allow the discriminator to learn local patterns and better discriminate real sketches from synthesized sketches. The discriminator D network structure is shown in Figure 3.

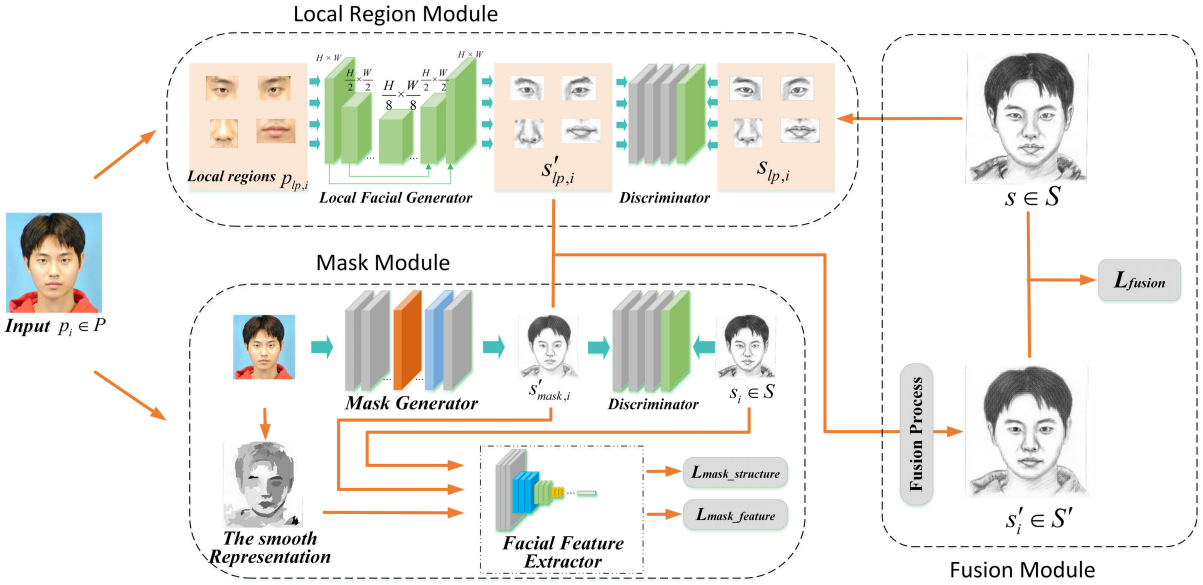

FIGURE 2. Face sketch synthesis framework based on MHGAN. The local region module divides the photo into multiple facial partial images: left eye, right eye, nose, and mouth, and input $p_{lp,i}$ into the generators G_{lp}, respectively. The mask module inputs the p_i to the mask generator G_{mask} to obtain the fake sketch mask $s'_{mask,i}$. The facial feature extractor separately calculates the MF loss of the real sketch and the fake sketch mask and MS loss of the smooth image of the real photo and the fake sketch. Outputs of five local nets are input to the fusion module and use with the landmarks to generate the final sketch s'_i.

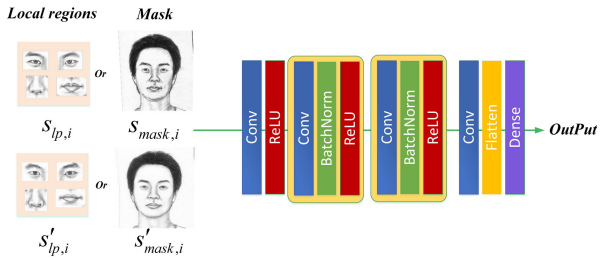

FIGURE 3. Structure of the discriminator network.

The loss function of the local region module is formulated as:

$$L_{local} = \gamma L_{local\_adv} + \delta L_{local\_l_1}, \tag{4}$$

where γ and δ are the hyperparameters which control the contribution of the L_{local_adv} and $L_{local\_l_1}$.

C. MASK MODULE

The mask module will provide the learning process of the mask hierarchical region, in this section $lp \in \{mask\}$. The input of the mask module is a real photo, which can ensure that the synthesized image does not lose high-level features and detailed features. This module includes a mask sketch synthesis network and a facial feature extractor.

1) MASK SKETCH SYNTHESIS NETWORK

The generator of the mask sketch synthesis network generates the coarse facial structure aims to preserve the position feature of the photos. This module uses the Earth-Mover (also known as Wasserstein-1) distance $W(P_{lp}, S'_{lp})$ proposed by Arjovsky *et al.* [28] to replace of the JS divergence in the standard GAN model. The objective function is as follow:

$$W(P_{lp}, S'_{lp}) = \max_{D_{lp} \in m} E_{p_{lp,i} \sim P_{lp}}[D_{lp}(p_{lp,i})] - E_{s'_{lp,i} \sim S'_{lp}}[D_{lp}(s'_{lp,i})], \tag{5}$$

where P_{lp} is the input photo distribution, S'_{lp} is the synthetic sketch sample distribution, $m = 1 - Lipschitz$. $W(P_{lp}, S'_{lp})$ is approximately the Wasserstein distance of the photo and the synthesized sketch. To ensure the discriminant network meets the 1-Lipschitz constraint, Gulrajani *et al.* [29] added a gradient penalty term in the objective function. The gradient penalty term is as follows:

$$\lambda E_{\hat{x} \sim P_{\hat{x}}}[(||\nabla_{\hat{x}} D_{lp}(\hat{x})||_2 - 1)^2], \tag{6}$$

where $P_{\hat{x}}$ is a random interpolation sample between the P_{lp} and S'_{lp}, $\hat{x}$ is the data randomly sampled between the real data and the generated data. Wasserstein distance can improve the mask sketch synthesis network's stability, ensure that important features such as the mask will not appear deformed and noise, and allow the local face region module to be more flexible. The adversarial loss in mask sketch synthesis is formulated as follow:

$$\begin{aligned} L_{mask\_adv}{}' = {} & E_{s_{lp,i} \sim S_{lp}}[D_{lp}(s_{lp,i})] \\ & - E_{s'_{lp,i} \sim S'_{lp}}[D_{lp}(s'_{lp,i})] \\ & + \lambda E_{\hat{s} \sim S_{\hat{s}}}[(||\nabla_{\hat{s}} D_{lp}(\hat{s})||_2 - 1)^2], \end{aligned} \tag{7}$$

where $S_{\hat{s}}$ is the random interpolation sample between the real sketch sample distribution S_{lp} and the synthesized sketch

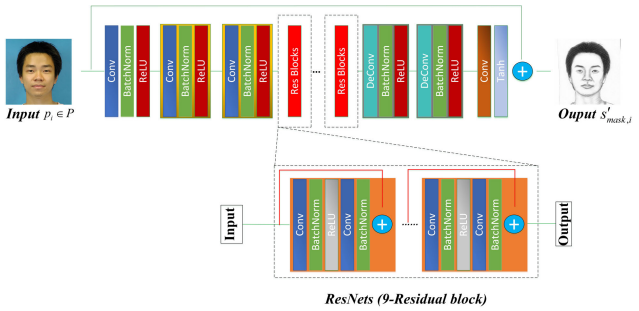

FIGURE 4. Structure of the residual network.

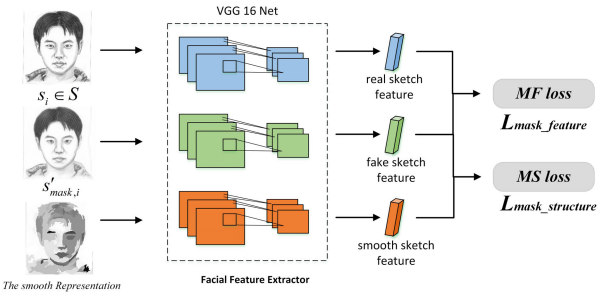

FIGURE 5. Extraction process of the facial feature extractor.

sample distribution S'_{lp}. To prevent the mask sketch synthesis model from no longer optimizing parameters due to loss reaches a premature equilibrium state, we add a weight control factor to the adversarial loss. The control factor can ensure that the proportion of adversarial loss is small at the beginning of training, and increase the weight later to ensure that the network does not converge quickly. It can improve the learning ability of the network training stage and improve the synthesis effect. Therefore, after adding the control factor the adversarial loss is rewrited as follow:

$$L_{mask\_adv} = \left(\frac{1+\omega n}{N}\right) L_{mask\_adv}{}', \tag{8}$$

where n is the current number of iterations, N is the total number of iterations, ω is the attenuation coefficient, and its value is fixed as 0.99.

The hierarchical generator G_{mask} in the mask module is constructed using a residual network. The structure of the generator G_{mask} is shown in Figure 4, and the input and output are shown in Eq. 9. The network of the discriminator has the same structure as the hierarchical discriminator in the local region module. As shown in Figure 3, the full mask sketch is input. The generator G_{mask} is a residual network with nine residual blocks [30]. The network structure includes a convolutional layer which stride is 1, two convolutional layers which stride is 2, nine residual units, two transposed convolutional layers, and a convolutional layer. The residual unit structure implemente through the shortcut connection and the linear forward process, mainly composed of the convolutional layer, batch Norm layer and ReLU activation function.

$$s'_{mask\,,i} = G_{lp}(p_{mask\,,i}). \tag{9}$$

2) FACIAL FEATURE EXTRACTOR

The extraction process of the facial feature extractor is shown in the Figure 5. It uses the VGG16 [31] network to extract features, and then MHGAN calculates two loss functions: MF loss and MS loss. MHAGN calculates the MF loss between the synthesized sketch and the real sketch, and the MS loss between the smooth image of the real photo and the synthesized sketch. The VGG16 network structure is built by repeatedly stacking a 3×3 convolution kernel and a 2×2 maximum pooling layer. The architecture dramatically reduces the number of network parameters by concatenating convolution kernels. Compared with the network layer constructed with a single convolution kernel, it has more nonlinear transformations and is more suitable for extracting facial features.

The purpose of the facial feature extractor is to compensate for the defects of the model-driven method to a certain extent, but the standard L_1 loss requirements are relatively strict, and the L_1 loss function is as follow:

$$L_1 = \sum_{i=1}^{w}\sum_{j=1}^{h}\left|s^{i,j} - G(p)^{i,j}\right|. \tag{10}$$

The L_1 loss calculates the distance in pixel-wise and directly obtains the real sketch's minimum absolute value and the synthesis sketch. It is difficult to obtain the facial details and texture features of the input sketch, and it loses the meaning of the facial feature extractor. Therefore, the proposed method uses the loss function of the facial feature extractor compares the sum of singular values between the two feature maps, which would improve the detailed features of the synthesized sketch:

$$L_{mask\_feature} = \frac{1}{wh}\left\|\phi(G_{lp}(p_{lp,\,i})) - \phi(s_{lp,\,i})\right\|_F{}^2, \tag{11}$$

where w, h is the dimensions of the feature map, $\phi(s_{lp,\,i})$ and $\phi(G_{lp}(p_{lp,\,i}))$ represent the feature matrix output from the real sketch and the fake sketch after feature extraction. $L_{mask\_feature}$ would bridge the feature gap in a latent space.

The smooth representation of the input photos can display sparse color blocks, global content, and clear boundaries. The input of the white box function is a gray-scale photo, and it will be divided into separate regions by using the felzenszwalb algorithm. Since the super-resolution algorithm only considers the similarity of pixels and ignore the semantic information, and use the average of pixel values to color each segmented region, we further introduce selective search [32] to merge the segmented regions and extract the sparse segmentation map. Finally, the white box function output a bright, smooth representation image. High-level features extracted through input the smooth image to the VGG16 network can impose spatial constraints between the synthesized result and the photo. The white box function is formulated in

Eq. 12 and 13.

$$S_{m,n} = (\theta_1 * \bar{S} + \theta_2 * \tilde{S})^{\mu}, \tag{12}$$

$$(\theta_1, \theta_2) = \begin{cases} (0,1) & \sigma(S) < \gamma_1 \\ (0.5, 0.5) & \gamma_1 < \sigma(S) < \gamma_2 \\ (1,0) & \gamma_2 < \sigma(S) \end{cases}, \tag{13}$$

where $S_{m,n}$ is a single pixel, $\bar{S}$ is the average value of the pixel, and $\tilde{S}$ is a similar pixel, and μ is a fixed parameter. According to the work of [33], γ_1 and γ_2 are divided into 20 and 40. We found that $\mu = 1.1$ is more suitable for gray-scale images and our method can generate good results. We define the function $F_{structure}$ is the process of the smooth image structure extraction and $L_{mask\_structure}$ is:

$$L_{mask\_structure} = \frac{1}{wh} \left\| \phi(G_{lp}(p_{lp,i})) - \phi(F_{structure}(p_{lp,\,i})) \right\|_F^{\,2}. \tag{14}$$

The model parameters of the facial feature extractor in the proposed method use the model parameters pre-trained by the VGG16 network on ImageNet [34].

The loss function of the mask module is:

$$L_{mask} = \alpha L_{mask\_adv} + \beta L_{mask\_feature} + \eta L_{mask\_structure}, \tag{15}$$

where α, β and η are the hyperparameters which control the contribution of the L_{mask_adv}, $L_{mask\_feature}$ and $L_{mask\_structure}$.

D. FUSION MODULE

Inspired by [35], the fusion module T uses the synthesized result $s'_{lp,\,i}$ of the local region module as the foreground image, and the synthesized result $s'_{mask,\,i}$ of the mask module as the background image. The final sketch s'_i is obtained by fusing the foreground image and background image. The proposed method chose to use the mixed gradients, which can fully select the texture features of the foreground image and the background image to achieve satisfactory results.

In the fusion module, where

$$s'_{lp,\,i} \in \{s'_{eye\_l,\,i}, s'_{eye\_r,\,i}, s'_{nose,\,i}, s'_{mouth,\,i}, s'_{mask,\,i}\}$$

in Figure 6 (a) and (b). Fusion module T can be formulated as:

$$s'_i = T(s'_{eye\_l,i}, s'_{eye\_r,i}, s'_{nose,i}, s'_{mouth,i}, s'_{mask,i}, landmark). \tag{16}$$

Furthermore, we use fake sketch s'_i and real sketch s_i to calculate the fusion loss to optimization the network, the fusion loss is defined as:

$$L_{fusion} = \frac{1}{wh} \left\| \phi(s_i) - \phi(s'_i) \right\|_F^{\,2} \tag{17}$$

(a)Local regions (b)Mask (c)Final Sketch

FIGURE 6. Fusion module result. (a) is the synthesized local region image, (b) is the synthesized background mask image, and (c) is the final sketch.

E. OPTIMIZATION THE SYNTHESIS NETWORK

The function Ψ can be formulated by solving the loss function expression is:

$$L_{total} = L_{local} + L_{mask} + \chi L_{fusion}. \tag{18}$$

where χ is the hyperparameter which control the contribution of the L_{fusion}. And algorithm 1 introduces the synthesis process of the network.

Algorithm 1 The Synthesis Process of the Network

Input: Real photo-sketch Z_{Data}; Base local region module Φ_0; Base mask module P_0; Base fusion module T;

1 for $i = 0 : N$;

2 Train Φ_i using Z_{lp} and Z'_{lp} in local region module and generate $s'_{lp,i}$, where $lp \in \{eye\_l, eye\_r, nose, mouth\}$. Compute L_{local} according to Eq. (4).;

3 Train P_i using Z_{lp} and Z'_{lp} in mask module and generate $s'_{lp,i}$, where $lp \in \{mask\}$. Compute L_{mask} according to Eq. (15).;

4 Input $s'_{lp,i}$ and landmark according to MHGAN's objection function T and generate fake sketch s'_i, where $lp \in \{mask\}$. Compute L_{fusion} according to Eq. (17).;

5 Fine-tune Φ_i and P_i according to Eq. (18);

Output: Network optimization loss L_{total};

IV. EXPERIMENT

A. DATASET

This paper's experiments are performed on two public face sketch datasets and photos on the internet. Two public datasets include the CUFS dataset and CUFSF dataset. The CUFS dataset consists of 606 photo-sketch pairs: 188 pairs from the Chinese University of Hong Kong (CUHK) [6] student dataset, 123 pairs from the AR dataset [36], and 295 pairs from the XM2VTS [37] dataset. The CUFSF dataset consists of 1194 sketch-photo pairs in total. Examples of these standard datasets are shown in Figure 7, 8, and Table 1 shows the training set and test set of these datasets divided.

FIGURE 7. Examples of the photo-image pairs from the CUFS dataset.

FIGURE 8. Examples of the photo-image pairs from the CUFSF dataset.

TABLE 1. Datasets division.

Dataset	Training pairs	Testing
CUHK Student	88	100
AR	80	43
XM2VTS	100	195
CUFSF	250	944

B. EXPERIMENTAL DETAILS

1) EXPERIMENTAL SETTINGS

The experiment of this paper is performed through PyTorch. [38]. All photo-sketch pairs in the datasets are geometrically aligned based on three points, i.e., two eye centers and the mouth center, then facial images above are cropped into 256×256. The input images size of the local region modules' hierarchical generators is 40×56 for eyes, 48×48 for nose, and 40×64 for the mouth. The input image size of the mask module is 256×256. Finally, the output image size of the fusion module is 256×256.

2) TRAINING DETAILS

We show how to determine each hyperparameter of the loss function and analyze the sensitivity of them in this subsection. We use a grid search method: the value of each hyperparameter was set to: 0.02, 0.1, 0.5, 1, 10, 15, 25, and we calculate the average FSIM of the CUFS dataset under different parameters, and select the parameter with the highest FSIM as the final parameter, which are $\delta = 25$, $\gamma = 1$, $\alpha = 1$, $\beta = 10$, $\eta = 0.02$ and $\chi = 0.02$.

In order to analyze the sensitivity of different loss hyperparameters to the overall loss, we fix the six hyperparameters of $\delta = 25$, $\gamma = 1$, $\alpha = 1$, $\beta = 10$, $\eta = 0.02$ and $\chi = 0.02$ in turn, and then adjust the remaining one hyperparameter to 0.02, 0.1, 0.5, 1, 10, 15, 25 and calculate the average FSIM of the CUFS dataset. When analyzing one parameter, we keep the others fixed. The experimental results are illustrated in Table 2.

TABLE 2. The FSIM of the different hyperparameters for overall loss.

	0.02	**0.1**	**0.5**	**1**	**10**	**15**	**25**
δ	0.7122	0.7287	0.7349	0.7422	0.7503	0.7521	**0.7549**
γ	0.7501	0.7519	0.7531	**0.7549**	0.7539	0.7524	0.7481
α	0.7524	0.7526	0.7538	**0.7549**	0.7533	0.7521	0.7489
η	**0.7549**	0.7540	0.7532	0.7512	0.7402	0.7316	0.7169
β	0.7356	0.7412	0.7473	0.7512	**0.7549**	0.7482	0.7480
χ	**0.7549**	0.7541	0.7538	0.7536	0.7522	0.7521	0.7444

It can be seen from the Table 2 that as the value of δ in the local region module increases, resulting in a continuous increase in the FSIM value, which indicates that δ has the greatest impact on the local region loss functions and the overall loss. Compared with δ, the influence of the parameter β of the mask module is the second important. At the same time, the loss parameters of γ and α that drive the image translation in the local region module and the mask module have less influence on the overall loss than δ and β. And it can be seen that the small-range fluctuations of χ and η have the least impact on the total loss, but if η is too large, it will have a greater impact on the mask sketch.

The learning rate and batch size are set to 0.0002 and 1, respectively. This paper chooses Adam [39] with β_1=0.5 and β_2=0.999 to optimize all modules, and the weight coefficients corresponding to each loss function are χ=η=0.02, $\gamma = 1$, $\delta = 25$, $\alpha = 1$, $\beta = 10$ from the above analysis and demonstration. The training process takes 300 epochs in total.

C. ABLATION STUDY

Our MHGAN combines several components for face sketch synthesis. Several ablation studies are conducted on the CUFS dataset to verify each component's contribution to the proposed method. Qualitative and quantitative evaluations are shown in Figure 9 and Table 3.

1. MHGAN without LRM (W/O LRM): The Local Region Module is removed. To verify the effectiveness of the LRM, the MHGAN only trains a single GAN as the generator model and the sketch generated by the mask module.

2. MHGAN without MF loss (W/O MF loss): The MF loss is removed in the mask module. The facial feature extractor only calculates the MS loss between the synthesized sketch and the smooth images.

3. MHGAN without MS loss (W/O MS loss): The MS loss is removed in the mask module. The facial feature extractor only calculates the MF loss between the synthesized sketch and the real sketch.

4. MHGAN without MG (W/O MG): The mixing gradient is removed in the fusion module. The fusion module uses normal guide field instead of the non-conservative guide field.

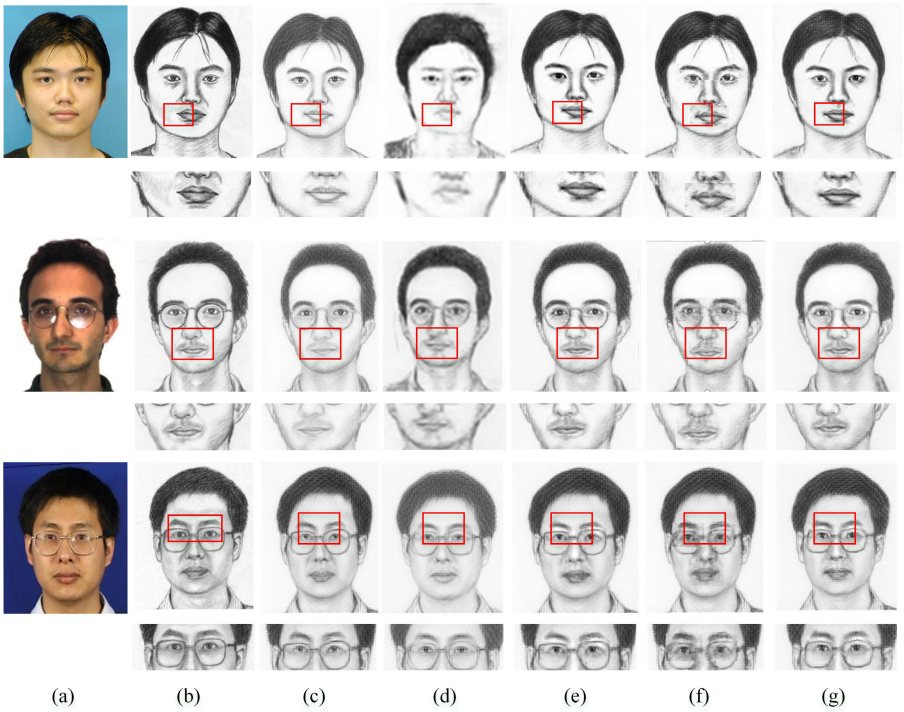

FIGURE 9. Synthesized sketches by our method for different settings on the CUFS dataset.(a) Original photos. (b) Original sketches. (c) MHGAN W/O LRM (Local Region Module). (d) MHGAN W/O MF loss. (e) MHGAN W/O MS loss. (f) MHGAN W/O MG (Mixing Gradient). (g) Ours. The eyebrows, eyes, nose, and mouth are refined. Our method can eliminate the noise in the synthetic sketch and without editing traces.

TABLE 3. Comparison of SSIM, FSIM and LPIPS to evaluate the quality of the synthesized sketch by our method for different settings on the CUFS dataset.

Dataset	Index	W/O LRM	W/O MF loss	W/O MS loss	W/O MG	Ours
CUHK	SSIM	0.5976	0.5200	0.6155	0.5195	**0.6201**
	FSIM	0.6704	0.6639	0.7155	0.6354	**0.7299**
	LPIPS	0.1977	0.2248	0.1946	0.2055	**0.1933**
AR	SSIM	0.7194	0.6228	0.7225	0.6043	**0.7333**
	FSIM	0.8042	0.7616	0.8334	0.7462	**0.8360**
	LPIPS	0.1669	0.1799	0.1705	0.1843	**0.1644**
XM2VTS	SSIM	0.5539	0.4877	0.5854	0.4581	**0.5863**
	FSIM	0.6794	0.7133	0.7146	0.6933	**0.7202**
	LPIPS	0.2330	0.2407	0.2229	0.2264	**0.2225**

The facial features of the face sketch contain a variety of drawing techniques and feature details. As shown in Figure 9 (c) and (g), the MHGAN W/O LRM cannot learn the detail of local facial region features well through a single GAN and appear blurry regions compared with our full method. The local region module of MHGAN is essential for capturing the texture details and styles of facial features.

As shown in the red boxes in Figure 9 (d) and (g), when the network converges, the results of MHGAN W/O MF loss (d) contain more blurry facial regions and lose some detail features, and without using MF loss will reduce the quality of synthetic sketches. As shown in the comparison of the red boxes in Figure 9 (e) and (g), the results of MHGAN W/O MS loss (e) contains some artifacts around the local regions, such as the mouth regions. The MS loss can further optimize the synthesized result, which makes the facial features' outline is more precise. The MF loss and MS loss in the mask module's facial feature extractor are crucial for generating a suitable mask.

Finally, we use the fusion module to maintain the boundary's consistency and obtain the fake sketch. As shown in Figure 9 (f) and (g), if non-conservative guide field is not used but other normal guide field in the fusion module, there will be noticeable edit marks at the boundarys of the sketch's facial regions. These traces significantly affect the subjective visual effect and are an apparent defect for the face sketch synthesis. It shows that the non-conservative

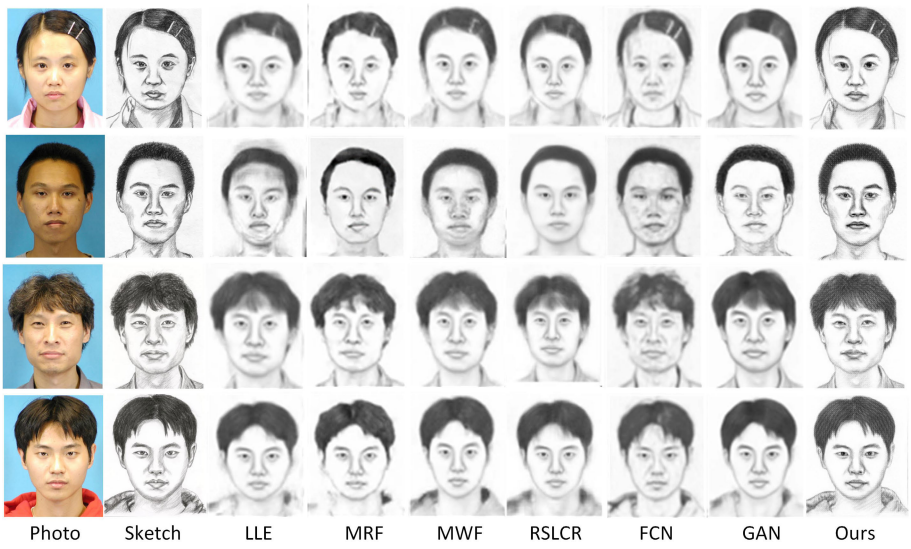

FIGURE 10. Face sketch synthesis results of different face sketch synthesis methods on the CUHK face dataset.

guide field is beneficial to improve the fusion effect of the synthesized sketch.

As shown in Table 3, the average SSIM, FSIM and LPIPS scores of MHGAN using all components are the best performance, which shows the effectiveness of each component in MHGAN.

D. COMPARISONS WITH STATE-OF-THE-ART METHODS

1) COMPARISON ON THE CUFS DATASET

We compared MHGAN with six state-of-the-art face sketch synthesis methods on the CUFS standard dataset, including the LLE [9] method, MRF [6] method, MWF [13] method, RSLCR [16] method, and FCN [19] method, and GAN [21] method. The comparison results are shown in Figure 10, 11, and 12.

As shown in Figure 10, 11, and 12, the synthesized results by LLE and MRF methods are relatively smooth and lose some common facial region structures. Such as the hairstyles at the third and fourth lines of AR, the facial contours at the second line of CUHK, and many artifacts at the first and third lines of XM2VTS. Although the MWF method can generate new candidate blocks by a weighted combination strategy but it cannot generate certain specific feature that only exists in the test sample, such as the glasses in the second and third rows of XM2VTS. The synthesized results by the RSLCR method are blurred at the facial contour and hairstyle. These data-driven methods only consider the similarity of images at the pixel level, so they cannot describe the facial characteristics well and leads to some problems such as blur and lack of texture features. Although the performance of the model-driven methods are better than the above methods and they can generate most facial region features, introducing some noise and reducing the sharpness of these results. For example, in the synthesized results by the FCN method, artifacts appear on the face, which affects the subjective visual effect. The synthesized sketches by the GAN method also loss the detail characteristics, such as headdresses. However, the quality of synthesized sketches by our method are the best. Its contain clear and sharp facial features, decorations, and other abundant low-level features. For example, the hairpins in the first line of CUHK, the hairstyles of AR, and the glasses in the first and second lines of XM2VTS have fine effective subjectively. From the synthesized results on the CUHK, AR, and XM2VTS datasets, we can found that the synthesized results can maintain complete facial structure and contours, mainly due to the MF loss and MS loss in the mask module of our method, which minimizes artifacts while ensuring an exact sketch effect. Furthermore, the Multi-Hierarchies division makes the synthesized result more clearer and sharper.

2) QUANTITATIVE EVALUATION ON THE CUFS DATASET

A quantitative analysis is performed to objectively prove the effectiveness of the MHGAN in the face sketch synthesis. Due to the lack of professional and objective evaluation methods for face sketches, we use traditional image quality assessment methods to evaluate the quality of synthesized sketch images. We utilize both the feature similarity index (FSIM) [40], structural similarity (SSIM) [41] and LPIPS [42] to evaluate the quality of synthesized sketches. The average FSIM, SSIM and LPIPS scores of the synthesized sketches are list in the Table 4. The higher FSIM, SSIM value, and the lower LPIPS value, the better the image quality. The numbers in bold in the table are the maximum values of each index. Figure 13 (a) and (b) show the FSIM and SSIM score statistics of all methods on the CUFS dataset. The horizontal axis shows the quality evaluation score, which ranges from 0 to 1. The vertical axis represents the percentage of synthesized

TABLE 4. Comparison of SSIM, FSIM and LPIPS to evaluate the quality of the synthesized sketch of different face sketch synthesis methods on the CUFS dataset.

Dataset	Index	LLE	MRF	MWF	RSLCR	FCN	GAN	Ours
CUHK	SSIM	0.5936	0.6027	0.6216	**0.6358**	0.6111	0.6297	0.6201
	FSIM	0.7134	0.7245	0.7235	0.7229	0.7075	0.7176	**0.7299**
	LPIPS	0.3349	0.3464	0.3253	0.3449	0.3786	0.2015	**0.1933**
AR	SSIM	0.6302	0.6063	0.6434	0.6518	0.6328	0.6500	**0.7333**
	FSIM	0.7538	0.7497	0.7600	0.7437	0.7288	0.7408	**0.8360**
	LPIPS	0.3485	0.3477	0.3384	0.3569	0.3941	0.1905	**0.1644**
XM2VTS	SSIM	0.4841	0.4668	0.4974	0.5168	0.4706	0.4899	**0.5863**
	FSIM	0.6867	0.6844	0.6947	0.6727	0.6787	0.6644	**0.7202**
	LPIPS	0.4028	0.3994	0.3939	0.4448	0.4719	0.2934	**0.2225**

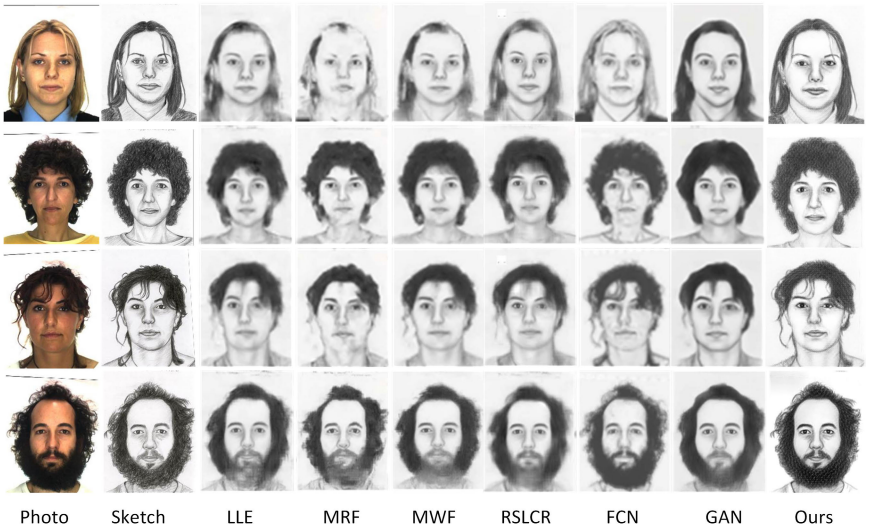

FIGURE 11. Face sketch synthesis results of different face sketch synthesis methods on the AR face dataset.

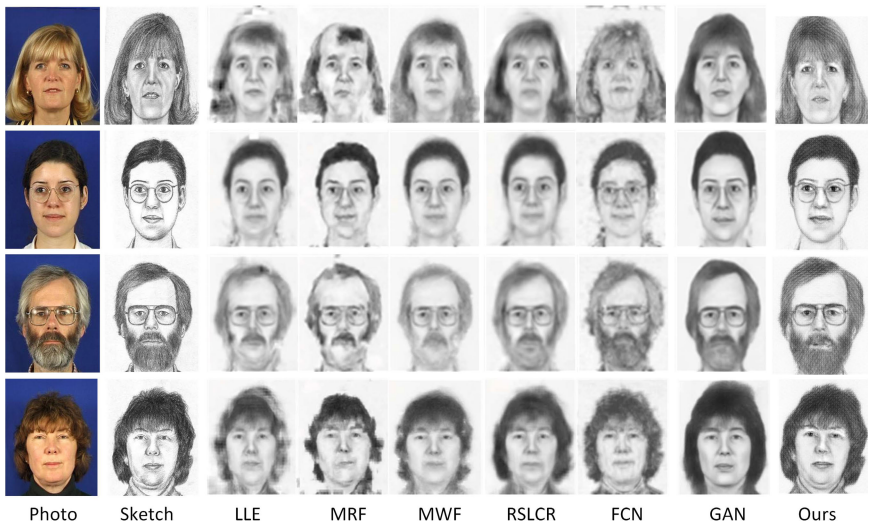

FIGURE 12. Face sketch synthesis results of different face sketch synthesis methods on the XM2VTS face dataset.

sketches, whose quality evaluation scores are larger than the score marked on the horizontal axis. Figure 13 (c) shows the LPIPS score statistics of all methods on the CUFS dataset. Unlike (a) and (b), the vertical axis in (c) represents the percentage of synthesized sketches, whose quality evaluation scores are lower than the score marked on the horizontal axis.

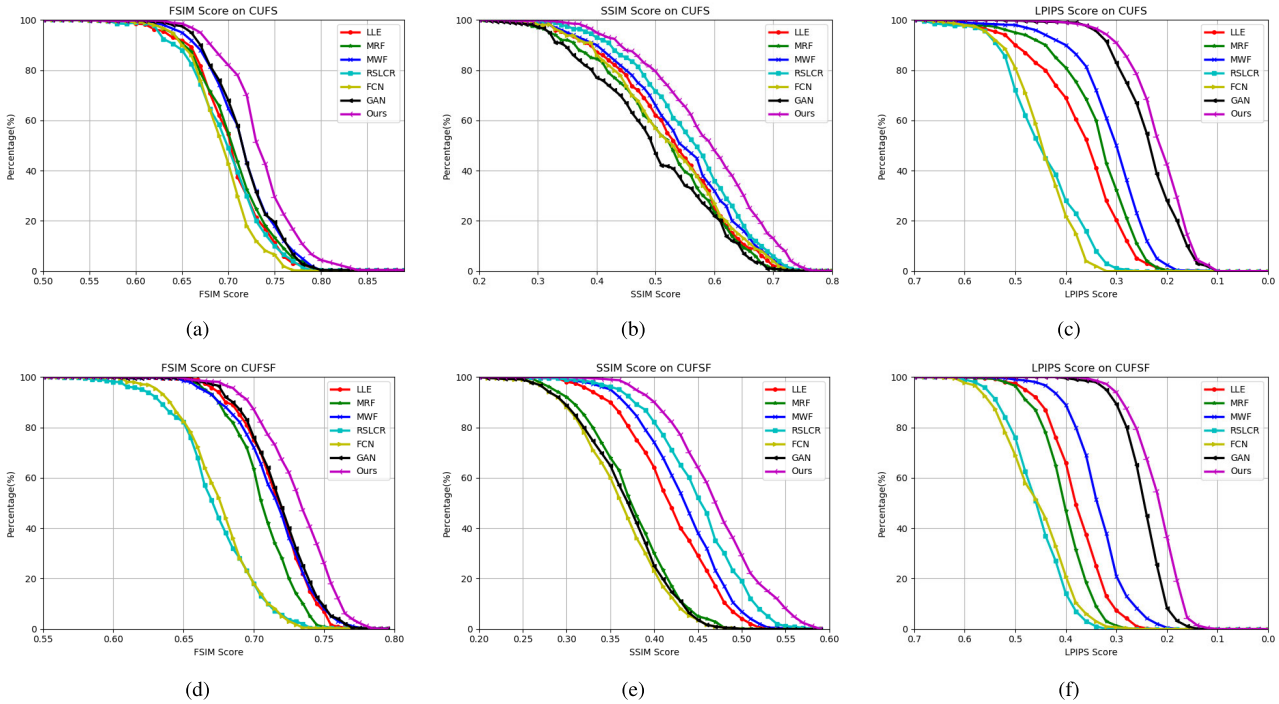

FIGURE 13. The performance of different methods on the CUFS dataset. (a) FSIM score on the CUFS dataset. (b) SSIM score on the CUFS dataset. (c) LPIPS Score on CUFS. (d) FSIM score on the CUFSF dataset. (e) SSIM score on the CUFSF dataset. (f) LPIPS Score on CUFSF.

TABLE 5. Comparison of SSIM, FSIM and LPIPS to evaluate the quality of the synthesized sketch of different face sketch synthesis methods on the CUFSF dataset.

Dataset	Index	LLE	MRF	MWF	RSLCR	FCN	GAN	Ours
CUFSF	SSIM	0.4172	0.5132	0.5393	0.5572	0.5213	0.4938	**0.6263**
	FSIM	0.7042	0.6956	0.7028	0.6648	0.6623	0.7059	**0.7301**
	LPIPS	0.3977	0.3748	0.4096	0.4529	0.5321	0.3102	**0.2403**

It can be seen from the curves that the MHGAN achieves higher performance than other methods on both SSIM, FSIM and LPIPS scores.

As shown in Table 4, the two average quality evaluation score of the MHGAN are higher than other methods on the CUFS dataset, which means that the generated sketches by our method is closer to the real sketch. Although the average SSIM score of the RSLCR method are higher than our method on the CUHK dataset, it have more blur regions comparing with the MHGAN in Fig. 10. In the work of [43], it was found that when the original sketch was used as the reference image, the FSIM score was closely related to human subjective evaluation. It can be seen from the evaluation score in Table 4 that the average FSIM score of the MHGAN on the CUFS dataset is higher than other methods. Furthermore, it can be seen from Table 4, among all the methods, our method achieves the best performance of LPIPS, which shows that the synthesized sketch by our method has the best perceptual quality and has the most similar texture detailed to the real sketch. Therefore, considering the subjective results and quality evaluation score, our method is very competitive and can generate high-quality sketches.

3) COMPARISON AND EVALUATION ON THE CUFSF DATASET

Compared with face photos, a corresponding sketch drawn by forensic experts in intelligent security applications is with shape exaggeration. To verify the robustness of our method on exaggerated images, we also conducted comparative experiments on the CUFSF dataset. The sketches drawn by artists in the CUFSF dataset are exaggerated in shape and expression. Table 5 shows the average SSIM, FSIM and LPIPS evaluation scores of the MHGAN and other methods on the CUFSF dataset. Figure 13 (d), (e) and (f) show the statistics of the FSIM, SSIM and LPIPS scores of all methods on the CUFSF dataset. In Figure. 14, the LLE, MRF, MWF, and RSLCR methods lose characteristic features, such as hairstreaks in the first and second lines. The FCN method has very complicated artifacts. Some facial components (such as mouth and eyes) are deformed in the GAN method, and some common facial structures are lost, such as facial contours. The comparison in Figure 14 shows that the proposed method's synthetic sketch is more vivid than other methods. The quantitative comparison in Table 5 and Figure 13 (d), (e) and (f) also shows that our method is superior to other methods.

FIGURE 14. Face sketch synthesis results of different face sketch synthesis methods on the CUFSF face dataset.

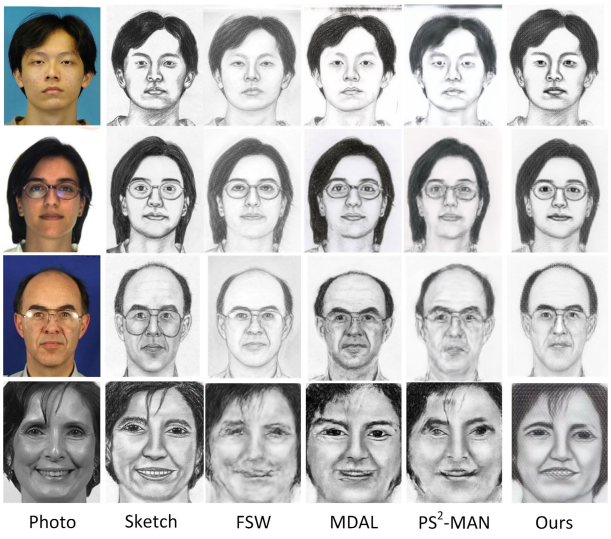

FIGURE 15. Face sketch synthesis results of different face sketch synthesis GAN-based methods on the CUFS and CUFSF datasets.

4) COMPARISON WITH GAN-BASED FACE SKETCH SYNTHESIS METHODS

We compare our method with the other three GAN-based face sketch synthesis methods. Figure 15 shows the face sketch synthesis results on the CUFS and CUFSF datasets, which are synthesized by the FSW [26] method, the MDAL [25] method, the PS^2-MAN [24] method, and our method. Table 6 shows the SSIM, FSIM and LPIPS evaluation results of four methods on two datasets. It can be seen that the results of the FSW method on the CUFSF dataset are smooth and blur. Although the MDAL and PS^2-MAN methods have been able to synthesize relatively good visual results on CUFS, the facial contours and partial structures are still blurred, and facial region features (such as the region around the eyes) are lost on CUFSF. The proposed method has sharp facial features, and it is better than other methods in the FSIM and LPIPS score.

TABLE 6. Comparison of SSIM, FSIM and LPIPS to evaluate the quality of the synthesized sketch of different face sketch synthesis GAN-based methods on the CUFS and CUFSF datasets.

Dataset	Index	FSW	PS^2-MAN	MDAL	Ours
CUFS	SSIM	0.5653	0.5998	**0.7507**	0.6574
	FSIM	0.7224	0.7438	0.7511	**0.7549**
	LPIPS	0.2912	0.2515	0.2249	**0.2064**
CUFSF	SSIM	0.4085	0.5818	0.5943	**0.6263**
	FSIM	0.6633	0.7057	0.6937	**0.7301**
	LPIPS	0.3222	0.3034	0.2804	**0.2403**

5) COMPARISON WITH GAN-BASED IMAGE-TO-IMAGE TRANSLATION METHODS

To furthermore illustrate the effectiveness of the MHGAN, it is compared with the other three Image-to-Image translation GAN-based methods. Figure 16 shows the synthesis results of different methods on the CUFS and CUFSF datasets, which are synthesized by the UNIT [44] method, the Cycle-GAN [23] method, the Dual-GAN [45] method, and our method. The synthesized sketches of the UNIT method are vaguer and reduce authenticity. Although the Cycle-GAN method can generate sketches with relatively good visual effects, but still has blurs and artifacts. The synthesized sketches by the Dual-GAN method has severe facial distortion. The proposed method can overcome blur and deformation, and the synthetic results are more realistic and clearer. Table 7 shows the SSIM, FSIM and LPIPS evaluation score of the four methods on the CUFS and CUFSF datasets. It can be seen that MHGAN is superior to the other

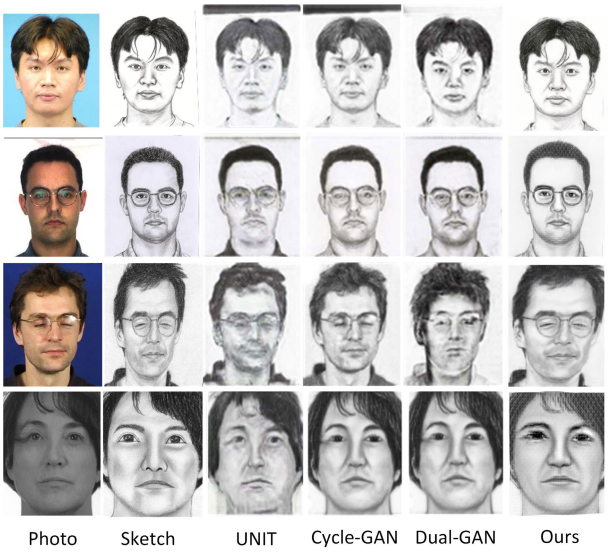

FIGURE 16. Face sketch synthesis results of different Image-to-Image translation GAN-based methods on the CUFS and CUFSF datasets.

TABLE 7. Comparison of SSIM, FSIM and LPIPS to evaluate the quality of the synthesized sketch of different Image-to-Image translation GAN-based methods on the CUFS and CUFSF datasets.

Dataset	Index	UNIT	Cycle-GAN	Dual-GAN	Ours
CUFS	SSIM	0.5804	0.5813	0.5581	**0.6574**
	FSIM	0.6997	0.7231	0.6898	**0.7549**
	LPIPS	0.3985	0.3586	0.3496	**0.2064**
CUFSF	SSIM	0.5711	0.5835	0.5732	**0.6263**
	FSIM	0.6759	0.7063	0.7057	**0.7301**
	LPIPS	0.4168	0.3883	0.3859	**0.2403**

three Image-to-Image translation GAN-based methods in the quality assessment.

6) FACE RECOGNITION

Face sketch recognition is usually used for quantitative evaluation of the face sketch synthesis [46], and it is also an important application of face sketch synthesis [47]. High-quality face sketch synthesis will have high-quality recognition accuracy. We employ the Null-space Linear Discriminant Analysis (NLDA) [48] for face recognition experiments to validate our method. In the recognition stage, all the photos in the dataset are first transformed to sketches by the face sketch synthesis method, and then match the input sketches with the synthesized sketches. For the CUFS dataset, we randomly select 150 synthesized sketch and corresponding real sketches to train the classifier, and the rest 188 synthesized sketches and the corresponding 188 original sketches as the testing set. For the CUFSF dataset, we randomly select 300 synthesized sketch and corresponding real sketches for training, and the rest 644 for testing. We randomly divided data on the CUFS dataset and the CUFSF dataset and repeated the face recognition experiment 20 times by NLDA. Figure 17 (a) and (b) show the face recognition rates

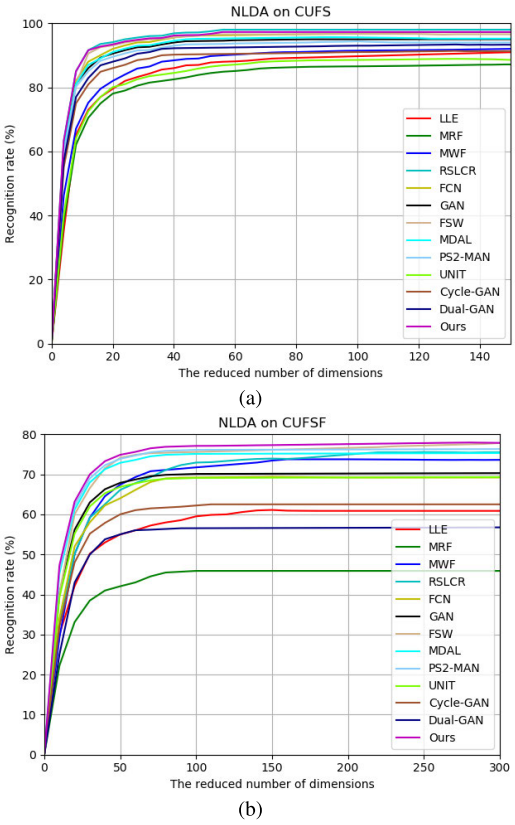

FIGURE 17. Face recognition accuracy on the (a) CUFS dataset and (b) CUFSF dataset.

against the numbers of dimensions on both CUFS and CUFSF datasets. Table 8 shows the best face recognition rate under a certain size. It can be concluded from Figure 17 and Table 8 that the recognition rate of the MHGAN model is the highest on the CUFSF dataset, and it is also very competitive compared with other methods on the CUFS dataset.

7) SKETCH SYNTHESIS ON THE REAL WORLD PHOTOS

The tested photos in the above experiment were all taken under specific conditions(such as lighting, background, facial expressions, etc). However, facial photos were taken by users of digital entertainment applications usually show different head poses and changing lighting environments. To further illustrate the applicability of our method in digital entertainment, we compare MHGAN with the MDAL and PS^2-MAN methods, which have better visual performance in the above experiments. Figure. 18 shows the synthesized sketches of these methods. The photos used in the test are all from the internet, and the training set is from the AR and CUHK datasets, containing 311 individuals. As shown in Figure. 18, although the synthesized sketches by the MDAL method have a rich texture, the eye structure of the sketch in the second line has been deformed. The synthesized sketches by the PS^2-MAN method in the third line lack important facial regions and texture of sketch. Compared with synthesized sketches by other methods, the synthesized sketches by our method in

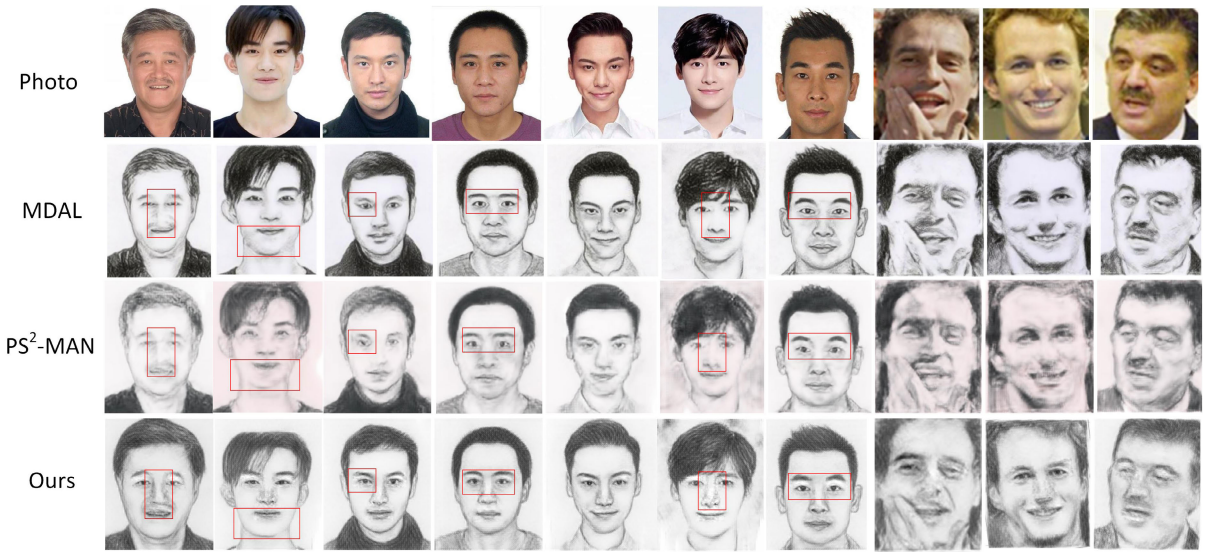

FIGURE 18. Different GAN-based face sketch synthesis results on internet photos. The second row are synthesized sketches by the MDAL method. The third row are synthesized sketches by the PS^2-MAN method. The fourth row are synthesized sketches by our method.

TABLE 8. NLDA face recognition accuracy based on sketches on the CUFS and CUFSF datasets.

Dataset	LLE	MRF	MWF	RSLCR	FCN	GAN	FSW	MDAL	PS^2-MAN	UNIT	Cycle-GAN	Dual-GAN	Ours
CUFS	0.9102 (143)	0.8718 (133)	0.9206 (139)	**0.9803 (87)**	0.9677 (138)	0.9491 (143)	0.9667 (89)	0.9567 (99)	0.9436 (136)	0.8891 (99)	0.9139 (140)	0.9339 (99)	0.9729 (146)
CUFSF	0.6090 (185)	0.4578 (272)	0.7376 (190)	0.7550 (295)	0.6938 (297)	0.7033 (141)	0.7781 (294)	0.7502 (146)	0.7629 (148)	0.7007 (143)	0.6251 (139)	0.5676 (114)	**0.7795 (239)**

the fourth line produces clear and more realistic, especially facial region details. This is due to the Multi-Hierarchies division of facial images in our method, which will not miss facial region features. The comparative experiments using real-world internet photos prove that our method has excellent performance and more robust applicability.

V. CONCLUSION

To address the problems that sketch local regions are easier to expose small artifacts and blur, we proposed a multi-hierarchies GAN-based face sketch synthesis method. A face photo is divided into multiple hierarchical structures and inputting them to the local region module and the mask module. The local region module can learn the detailed features of different local regions of the face and generate local region sketches. The mask module generates a coarse facial structure of a sketch. Finally, the local region sketches and the mask sketch are input to the fusion module and generate the fake sketch. Through experiments on the different datasets illustrate that the proposed method can synthesis sketch with good details and fine facial textures. Some quantitative evaluation shows that our method has achieved better performance than the state-of-the-art methods and is robust. In the future, we will modify the encoder or fusion module to improve our method's synthesis speed. We also intend to enhance our method's practical application in the wild.

REFERENCES

[1] M. Zhang, N. Wang, Y. Li, and X. Gao, "Bionic face sketch generator," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2701–2714, Jun. 2020.
[2] M. Zhang, J. Zhang, Y. Chi, Y. Li, N. Wang, and X. Gao, "Cross-domain face sketch synthesis," *IEEE Access*, vol. 7, pp. 98866–98874, 2019.
[3] K.-K. Huang, D.-Q. Dai, C.-X. Ren, and Z.-R. Lai, "Learning kernel extended dictionary for face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 5, pp. 1082–1094, May 2017.
[4] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "Transductive face sketch-photo synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 9, pp. 1364–1376, Sep. 2013.
[5] M. Zhang, R. Wang, X. Gao, J. Li, and D. Tao, "Dual-transfer face sketch–photo synthesis," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 642–657, Feb. 2019.
[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] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. CVPR*, Jun. 2011, pp. 513–520.
[8] X. Tang and X. Wang, "Face photo recognition using sketch," in *Proc. Int. Conf. Image Process.*, vol. 1, 2002, p. 1.
[9] S. T. Roweis, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.
[10] P. Cheng, J. Wang, S. He, X. Luan, and F. Liu, "Observer-based asynchronous fault detection for conic-type nonlinear jumping systems and its application to separately excited DC motor," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 3, pp. 951–962, Mar. 2020.
[11] P. Cheng, S. He, J. Cheng, X. Luan, and F. Liu, "Asynchronous output feedback control for a class of conic-type nonlinear hidden Markov jump systems within a finite-time interval," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Mar. 25, 2020, doi: 10.1109/TSMC.2020.2980312.
[12] P. Cheng and S. He, "Observer-based finite-time asynchronous control for a class of hidden Markov jumping systems with conic-type nonlinearities," *IET Control Theory Appl.*, vol. 14, no. 2, pp. 244–252, Jan. 2020.

[13] 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.
[14] X. Gao, N. Wang, D. Tao, and X. Li, "Face sketch–photo synthesis and retrieval using sparse representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 8, pp. 1213–1226, Aug. 2012.
[15] S. Zhang, X. Gao, N. Wang, J. Li, and M. Zhang, "Face sketch synthesis via sparse representation-based greedy search," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2466–2477, Aug. 2015.
[16] N. Wang, X. Gao, and J. Li, "Random sampling for fast face sketch synthesis," *Pattern Recognit.*, vol. 76, pp. 215–227, Apr. 2018.
[17] L. Chang, M. Zhou, X. Deng, Z. Wu, and Y. Han, "Face sketch synthesis via multivariate output regression," in *Proc. Int. Conf. Hum.-Comput. Interact.* Berlin, Germany: Springer, 2011, pp. 555–561.
[18] M. Zhu and N. Wang, "A simple and fast method for face sketch synthesis," in *Proc. Int. Conf. Internet Multimedia Comput. Service (ICIMCS)*, 2016, pp. 168–171.
[19] 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)*, 2015, pp. 627–634.
[20] B. Sheng, P. Li, C. Gao, and K.-L. Ma, "Deep neural representation guided face sketch synthesis," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 12, pp. 3216–3230, Dec. 2019.
[21] 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.
[22] M. Zhang, N. Wang, Y. Li, R. Wang, and X. Gao, "Face sketch synthesis from coarse to fine," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI), 30th Innov. Appl. Artif. Intell. (IAAI), 8th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, S. A. McIlraith and K. Q. Weinberger, Eds. New Orleans, LA, USA: AAAI Press, Feb. 2018, pp. 7558–7565.
[23] 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.
[24] L. Wang, V. Sindagi, and V. Patel, "High-quality facial photo-sketch synthesis using multi-adversarial networks," in *Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*, May 2018, pp. 83–90.
[25] 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.
[26] C. Chen, W. Liu, X. Tan, and K.-Y. K. Wong, "Semi-supervised learning for face sketch synthesis in the wild," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 216–231.
[27] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[28] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: http://arxiv.org/abs/1701.07875
[29] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[32] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.
[33] X. Wang and J. Yu, "Learning to cartoonize using white-box cartoon representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8090–8099.
[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[35] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *Proc. ACM SIGGRAPH Papers*, 2003, pp. 313–318.
[36] A. M. Martinez, "The ar face database," CVC, New Delhi, India, Tech. Rep. 24, 1998.
[37] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio Video-Based Biometric Person Authentication*, vol. 964, 1999, pp. 965–966.
[38] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. NeurIPS Workshop*, 2017.
[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[40] 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.
[41] 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.
[42] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[43] N. Wang, "Heterogeneous facial image synthesis and its applications," (in Chinese), Ph.D. dissertation, Xidian Univ., Xi'an, China, 2014.
[44] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 700–708.
[45] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2849–2857.
[46] Y. Song, L. Bao, Q. Yang, and M.-H. Yang, "Real-time exemplar-based face sketch synthesis," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 800–813.
[47] Y. Fang, W. Deng, J. Du, and J. Hu, "Identity-aware CycleGAN for face photo-sketch synthesis and recognition," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107249.
[48] 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.

KANGNING DU received the B.Sc. degree in telecommunication engineering from Beijing Information Science and Technology University, in 2011, and the Ph.D. degree in communication and information system from the Institute of Electronics, Chinese Academy of Sciences, in 2016. He is currently a Teacher of electronic engineering with Beijing Information Science and Technology University. His research interests include radar signal processing and image understanding and recognition.

HUAQIANG ZHOU received the B.Sc. degree from Beijing Information Science and Technology University, in 2018, where he is currently pursuing the M.Sc. degree. His research interests include machine learning and computer vision.

LIN CAO received the B.Eng. degree in telecommunication engineering from Northeastern University, China, in 1999, and the Ph.D. degree in signal and information processing from the Institute of Electronics, Chinese Academy of Sciences, in 2005. He is currently a Professor with the Department of Electronic Engineering, Beijing Information Science and Technology University (BISTU). He teaches courses on digital signal processing, digital image processing, and soft design fundamentals. He is the Dean of the School of Information and Communication Engineering and the Deputy Director of the Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, BISTU. He is a member of China Education Society of Electronics. He has published over 40 articles on image processing and pattern recognition. His research interests include radar signal processing and image understanding and recognition.

YANAN GUO received the B.Sc. degree from Hubei Polytechnic University, in 2014, and the Ph.D. degree from Yunnan University, in 2019. She is currently a Teacher of electronic engineering with Beijing Information Science and Technology University. Her research interests include machine learning and computer vision.

TAO WANG received the Ph.D. degree from the School of Electronic and Information Engineering, Beihang University, in 2019. He is currently a Teacher of electronic engineering with Beijing Information Science and Technology University. His research interests include radar signal processing, data fusion, and target localization and tracking in intelligent transportation systems or vehicle intelligent assistance systems.