Neurocomputing 365 (2019) 113–124

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

FCN based preprocessing for exemplar-based face sketch synthesis

Dan Lu[a], Zhenxue Chen[a,b,*], Q.M. Jonathan Wu[c], Xuetao Zhang[a]

[a] School of Control Science and Engineering, Shandong University, Jinan, 250061, PR China
[b] Shenzhen Research Institute of Shandong University, Shandong University, Shenzhen 518057, China
[c] Department of Electrical and Computer Engineering, University of Windsor, Windsor N9B 3P4, Canada

ARTICLE INFO

Article history:
Received 5 November 2018
Revised 24 May 2019
Accepted 16 July 2019
Available online 18 July 2019

Communicated by Prof. Zidong Wang

Keywords:
Face sketch synthesis
Exemplar-based
Fully convolutional network
Preprocessing

ABSTRACT

Most of the current exemplar-based face sketch synthesis approaches directly synthesize face sketches from face photos. However, due to the great difference between face photos and sketches, as well as the cluttered backgrounds in photo images, there tends to be some noise, deformation and missing parts on the synthesized face sketches by most of the exemplar-based methods. Besides, most exemplar-based methods exist a common problem: they only produce satisfactory results when training and test samples originate from the same dataset. To address these issues, in this paper we propose a simple but effective method which consists of two stages: the preprocessing stage and the sketch synthesis stage. In the preprocessing stage, we first design a fully convolutional neural network for preprocessing (pFCN). To fit the preprocessing task, the pFCN is trained by an L1 based total loss function, which is simple yet could enhance the facial features. Then the full-size photo is fed to the well-trained pFCN to generate the feature map, which we call a *semi-sketch* since it bridges the discrepancy between photo and sketch. At the sketch synthesis stage, the semi-sketches and an existing exemplar-based method are employed to synthesize the final sketches. Extensive experiments on public face sketch datasets verify that the proposed two-stage method improves the sketch synthesis quality of the state-of-the-art exemplar-based methods in terms of both recognition accuracy and perceptual quality. In addition, the experiments on cross-dataset indicate that the proposed method provides a new means for strengthening the generalization ability of the exemplar-based method.

1. Introduction

A sketch is a quick, rough drawing that shows the main features of an object or scene. Synthesizing face sketches from photos, an important branch of heterogeneous image transformation (HIT), has been widely used in both law enforcement and digital entertainment [4–6]. Surveillance cameras have been widely used in law enforcement, as an important tool for maintaining public order. However, in many criminal cases, surveillance cameras only provide limited information about suspects, which is not enough for normal face recognition methods. Under these circumstances, a sketch drawn by the artist based on the recollection of an eyewitness and the clues from video surveillance is often considered the best substitute for suspect identification [7,8]. Synthesizing face sketches from the photo database, and then taking the sketch drawn by the artist as the probe to retrieve from the synthesized sketches can help the police quickly identify the suspect. In digital entertainment, people like to display their photos in an artistic style, and automatic sketch portrait generation could help them, being faster and more convenient. Therefore, the synthesized sketches need to meet two requirements: (1) easy to identify; (2) vivid and delicate in visual perception.

Significant progress has been made in face sketch synthesis, over the past decade. Thanks to the rapid development of machine learning and deep learning, the face sketch synthesis methods have become more diverse, and the quality of the synthesized sketch has also significantly improved. Exemplar-based methods play an important role among the various face sketch synthesis methods. This is due to their success in detecting and exploiting patch correspondences within a training database or calculating optimized dictionaries allowing for highly sparse data representation [9].

Given a photo patch, exemplar-based methods will first search neighbor patches in the training photo set, and then their corresponding sketch patches are linearly combined to synthesize the target sketch. In most of the exemplar-based methods, the combination weights are computed based on the assumption that if two photo patches are similar, then their corresponding sketch patches are also similar [3]. However, this assumption is not always true;

* Corresponding author.
E-mail addresses: chenzhenxue@sdu.edu.cn (Z. Chen), jwu@uwindsor.ca (Q.M.J. Wu).

https://doi.org/10.1016/j.neucom.2019.07.008
0925-2312/

Fig. 1. Face sketches synthesized by three existing exemplar-based methods and the proposed two-stage methods. (b) LLE [1], (c) MRF [2], (d) RSLCR [3]; (f)–(h) are from our proposed two-stage method.

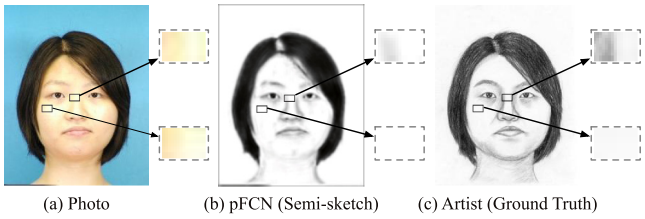


Fig. 2. Even if the two photo patches are very similar, their corresponding sketch patches might be very different; however, introduction of the semi-sketch alleviates this contradiction.

sometimes even when the photo patches are very similar, their corresponding sketch patches may be different [2], as shown in Fig. 2. In addition, the results of the exemplar-based methods are greatly influenced by the distribution of training samples, since they are reconstructed from the training samples. Therefore, if the distribution of the test sample is different from the training samples, it is hard for these methods to synthesize satisfactory results. As shown in Fig. 1, the photo (see Fig. 1(a)) is a little darker than the training samples, and there are some unsatisfactory parts of the synthesized results, such as the noise in the background and skin area, the deformation of the nose and mouth, and obvious gridding problems (see Fig. 1(b)–(d)). These problems will be more serious when the test samples and training samples are not from the same dataset. Thus, using an appropriate preprocessing method to reduce the distribution difference between test samples and training samples and make the above-mentioned assumption stronger is an effective way to improve the synthesis quality of the exemplar-based methods.

Recently, several works [10,11] have successfully exploited convolutional neural network (CNN) to synthesize face sketches from photos. Although some of them do not perform better than some exemplar-based methods in terms of recognition accuracy, they are able to retain the structure and content of the photos well, since they can generate sketches directly and globally rather than using training sketch patches to synthesize sketches. Inspired by this, rather than the commonly used exemplar-based strategy where sketches are directly synthesized from photos, this paper proposes a two-stage method, consisting of a preprocessing stage and a sketch synthesis stage. In the preprocessing stage, we take training photos as inputs and their corresponding sketches as labels to train a fully convolutional neural network. Since this fully convolutional neural network is used for preprocessing, we denote it as pFCN. Then the training photos and test photos are fed

Fig. 3. Comparison between shading sketch and profile sketch. Image (c) is from [17].

to the well-trained pFCN to get training semi-sketches and test semi-sketches. Semi-sketches not only retain the structure and content information of the photos but also have some characteristics of the sketches, such as the white background. At this stage, both training photos and test photos are transformed into semi-sketches and their distribution differences are greatly diminished. In the sketch synthesis stage, we directly employ an existing exemplar-based method to synthesize the final sketches, but for the inputs we replace the training photos and test photos with training semi-sketches and test semi-sketches. Specifically, we design a simpler fully convolutional neural network (pFCN) than [10,11], because the texture learning is not crucial in the preprocessing stage. To learn more details about the key facial features, a total loss function which includes a global loss function and two local loss functions are used to train the pFCN.

The contributions of this work are mainly three-fold.

First, we propose a two-stage (a preprocessing stage and a sketch synthesis stage) method for face sketch synthesis. Specifically, the proposed two-stage method takes a fully convolutional neural network as the preprocessing of the exemplar-based method.

Second, we design a simpler neural network architecture inspired by Zhang et al. [11], termed pFCN. Then a simple yet effective loss function is designed to focus training on learning more structural and content information.

Third, detailed experiments are conducted on the CUFS database [2] to demonstrate the improvement in the synthesis quality. Besides, the experimental results on cross-dataset show that the proposed preprocessing can improve the generalization ability of the existing exemplar-based method.

2. Related work

In this section, previous works on exemplar-based face sketch synthesis and dense predictions via CNNs are reviewed.

2.1. Exemplar-based face sketch synthesis methods

In recent decades, researchers have made great efforts in the field of sketch face synthesis and achieved remarkable results. Based on the previous studies [10,12,13], exemplar-based sketch face synthesis can be roughly divided into two categories: profile sketch synthesis [14–16] and shading sketch synthesis [3,12,13]. As we can see from Fig. 3, profile sketches are more like line drawings. Compared with profile sketches, the shading sketches can not only use lines to reflect the overall profiles but also capture the textural parts via shading [10]. Therefore, shading sketches are more expressive than profile sketches. This paper focuses on shading sketch synthesis.

Tang and Wang [18] used a separate eigen-transformation algorithm (ET) to synthesize a face sketch from a photo. This algorithm assumed that the photo has a linear correspondence

with its corresponding sketch if their shape and texture were treated independently. However, this assumption may be too strong for all face photos and sketches, especially when considering the hair area. Inspired by the manifold learning method called locally linear embedding (LLE) [19,20], Liu et al. [1] proposed a face sketch synthesis method that works on image patch level. Different from the global linear assumption in [1,18] was based on the idea of locally linear approximating global nonlinear. This method first divided the photo and sketch into overlapping image patches in the same way. Then for a test photo patch, K nearest neighbors were selected from training photo patches. Then the K sketch patches which corresponded to the K nearest photo patches were used to reconstruct the target sketch in the weighted linear combination way. The reconstruction weights were calculated in the spirit of locally linear embedding. However, due to the target sketch patches are synthesized independently at a fixed scale, the face shape cannot be well learned. Wang and Tang [2] proposed a method that synthesizes target sketch patches at different scales by using a multiscale Markov random fields (MRF) model, which had a profound impact on subsequent research. This method only finds one most appropriate photo patch from the training set and uses its corresponding sketch patch to estimate the target sketch patch. Therefore, it is difficult to deal with new patches that have never appeared in the training set. In addition, the MRF's optimization is NP-hard. To address these problems, Zhou et al. [12] proposed a method named Markov weight field (MWF), which introduced the linear combination into MRF and considered the dependency constraint between adjacent synthesized sketch patches. They formulated their model into a convex quadratic programming (QP) problem and proposed a cascade decomposition method (CDM) to solve this QP problem. Zhang et al. [21] also proposed a method based on MRF, which focuses on improving the robustness to lighting and pose variations. Similarly, [22] also made a significant contribution to coping with the problem of lighting variation in face sketch synthesis by using a preprocessing method named bidirectional luminance remapping (BLR). Unlike MRF and MWF, which use a single representation to measure the similarity between two image patches, Peng et al. [23] used a combination of multiple representations and obtained impressive results. To reduce the time consumption while ensuring the quality of the synthesis, Wang et al. [3] presented a face sketch synthesis framework based on random sampling and locality constraint.

2.2. Dense predictions via CNNs

Convolutional neural networks (CNNs) are a kind of neural network model, whose architectures usually have three types of layers: convolutional layer, pooling layer, and fully-connected layer. Nowadays, CNNs have produced impressive results in many traditional computer vision tasks, such as object detection, localization, semantic segmentation, classification and recognition [24–28]. Specifically, dense prediction, one of the traditional areas, has also achieved rapid development due to the introduction of CNNs. Dense prediction refers to per-pixel prediction from one or more input images. Long et al. [29] transformed the fully-connected layers in a classification net into convolution layers to build a fully convolutional network. Their work demonstrated that fully convolutional network is a desirable choice for solving dense prediction problems in per-pixel tasks like semantic segmentation, due to its ability to take arbitrarily sized inputs and return spatial outputs. Sermanet et al. [25] proposed an integrated framework based on CNN, which is used for classification, localization and detection. Liu et al. [30] designed a deep convolutional neural field model to solve the problem of depth estimation from a single image and achieved a state-of-the-art result without using geometric priors. The highlight of [30] is that they incorporated the optimization problem in a continuous conditional random field (CRF) into a deep CNN framework. Dong [31] developed a three-layer convolutional neural network to learn an end-to-end mapping between the low-resolution image and its corresponding high-resolution image.

Inspired by Dong et al. [31], Zhang et al. [11] designed a six-layer fully convolutional neural network (FCN), which takes photo image (RGB channels and two channels of the corresponding coordinate (i, j)) as input and outputs target sketch image. The sketches synthesized by FCN greatly reduce the discrepancy between photos and sketches, but they have blurry contours due to the mean square error metric (MSE) in the training loss. Zhang et al. [10] also proposed a CNN-based method and trained end-to-end. To enhance the texture of the hair area, they designed a two-branch fully convolutional neural network (BFCN), which generates structural and textural representations. Although the synthesized sketches achieve impressive results in sketch-based face recognition, the texture of these sketches does not look natural.

3. Photo–sketch synthesis

Most of the exemplar-based methods assume that similar photo images have similar sketch images [10]. However, due to the texture and shape discrepancy between photos and sketches, sometimes even though two photo patches are very similar, their corresponding sketch patches might be very different, as shown in Fig. 2. Besides, the cluttered background may cause some noise in the synthesized sketches. Moreover, the requirement of distribution similarity between training samples and test samples leads to the exemplar-based method cannot synthesize satisfactory results on cross-database [10]. Therefore, using preprocessing to reduce the gap between photos and sketches, simplify the background and reduce the distribution differences between training samples and test samples can improve the synthetic quality of the existing exemplar-based methods.

In this paper, we propose a method consisting of two stages: the preprocessing stage and the synthesis stage. In the preprocessing stage, a fully convolutional neural network (pFCN) is used to generate the *semi-sketch*, which is the transition state of the photo and sketch. In the synthesis stage, the semi-sketches are used in the existing exemplar-based methods to synthesize the target sketch. A graphical pipeline of the proposed two-stage method is shown in Fig. 4.

In the following sections, we introduce the preprocessing stage and the synthesis stage, respectively. For the preprocessing stage, we describe the details of the proposed pFCN framework. For the synthesis stage, the random sampling with locality constraint for face sketch synthesis method (RSLCR) [3] is used as an example of the exemplar-based method to introduce how to use the semi-sketches to synthesize the target sketch.

3.1. Semi-sketch generation via pFCN

The main task in the preprocessing stage is as follows. Given a face photo image $\mathcal{P}$, we would like to generate a face semi-sketch image $\mathcal{X}$, which has details of the facial features (like eyes, nose and mouth), the main structure of the face (such as the distance between the eyes) and simplified background. For each face photo image $\mathcal{P}$, the generated face semi-sketch image needs to have a similar distribution. In this paper, we design a fully convolutional neural network to accomplish this task. The final task is synthesizing a target sketch which is as close as possible to the hand-drawn sketch, and the hand-drawn sketch (ground truth) also meets the requirements of the semi-sketches as to content, structure and distribution. Considering these, we take the face photo image as the input and its corresponding hand-drawn sketch as the label to train the fully convolutional neural network.

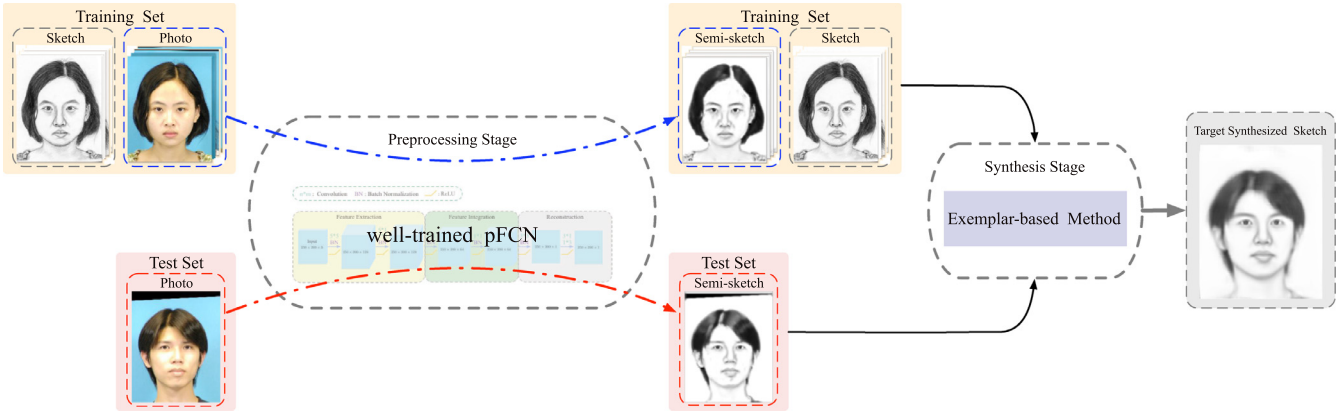


Fig. 4. Illustration of the proposed two-stage method. The architecture of pFCN is illustrated in Fig. 5.

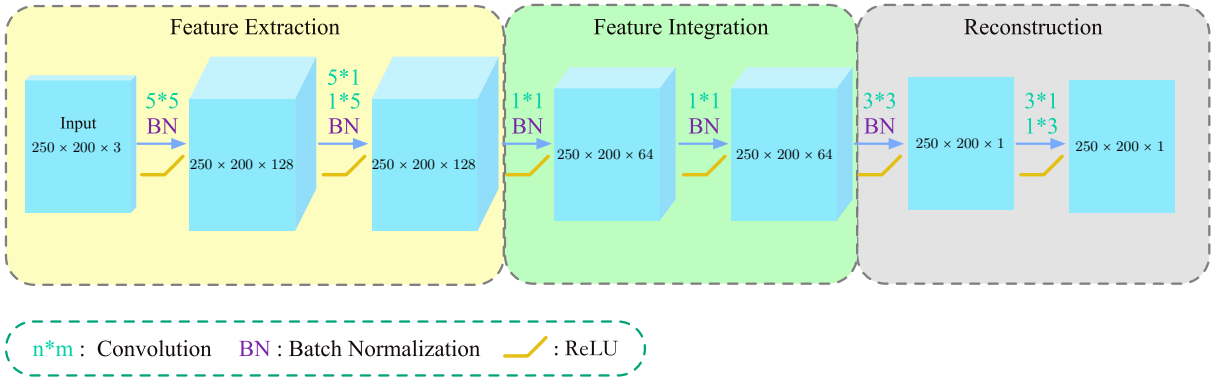


Fig. 5. The architecture of the proposed pFCN. This model takes a photo image as input and outputs a semi-sketch of the same size.

In particular, the previous work in [11] has successfully applied a six-layer fully convolutional neural network (FCN) to synthesize sketches. And these synthesized sketches present the main content of the photos and some texture of the sketches. Inspired by FCN [11], we design an eight-layer fully convolutional neural network for preprocessing (namely pFCN).

3.1.1. Network architecture

The architecture of pFCN is shown in Fig. 5. Unlike the FCN which uses the position prior and photo image as the input, pFCN directly takes the photo image as the input. Moreover, pFCN has 8 convolutional layers, since we decomposed the latter of the two 5×5 (3×3) convolution layers to 5×1 and 1×5 (3×1 and 1×3). This operation can reduce the number of parameters, which is good for reducing overfitting in the training process when the scale of the training set is small. Our network adapts rectified linear unit (ReLU) as activation function. Besides, we add batch normalization before each ReLU except the last layer.

3.1.2. Loss function

Since the texture can be added in the synthesis stage, the process of generating the semi-sketch needs to pay more attention to learning the content of face photo rather than the texture of the face sketch. To fit our task, this paper applies L1-norm between the generated semi-sketch $\mathcal{X}$ and hand-drawn sketch $\mathcal{S}$ as the loss function. L1-norm will preserve more details than L2-norm [32], and these details play an important role in the following sketch synthesis stage. The global loss function can be formulated as:

$$\mathcal{L}_{global} = \frac{1}{N}\sum_{i=1}^{N} \|\mathcal{X}_i - \mathcal{S}_i\|_1 \tag{1}$$

where N is the number of training samples.

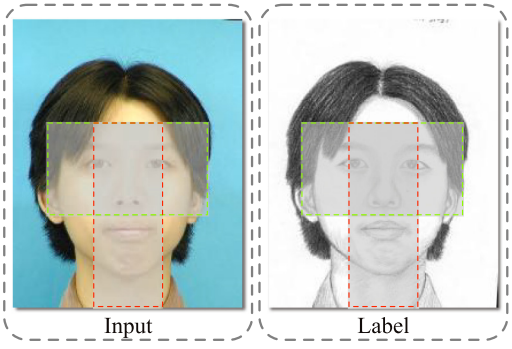


Fig. 6. Illustration of the calculation area of the local loss functions.

When drawing a face sketch image, the facial features (eyebrows, eyes, nose and mouth) will take up most of the time, although the area they occupy in the overall image is small. To obtain a high-quality result, some methods imitated the hand-drawn process to strengthen the supervision of the facial features, such as CA-GAN [33]. CA-GAN employed the face parsing method proposed by Liu et al. [34] to decompose the facial features and strengthen supervision over them. In addition to the above-mentioned facial features, we found that the smile folds and under-eye bags have also been stressed in the hand-drawn sketch (ground truth), which may provide useful help for identification. However, it is hard to find a face parsing method to decompose these two components.

This paper adopted a simple but effective method to strengthen the supervision of these facial features. As shown in Fig. 6, there are two gray rectangles in every face image. The edge of one rectangle is a green dashed line, and the edge of the other is the red

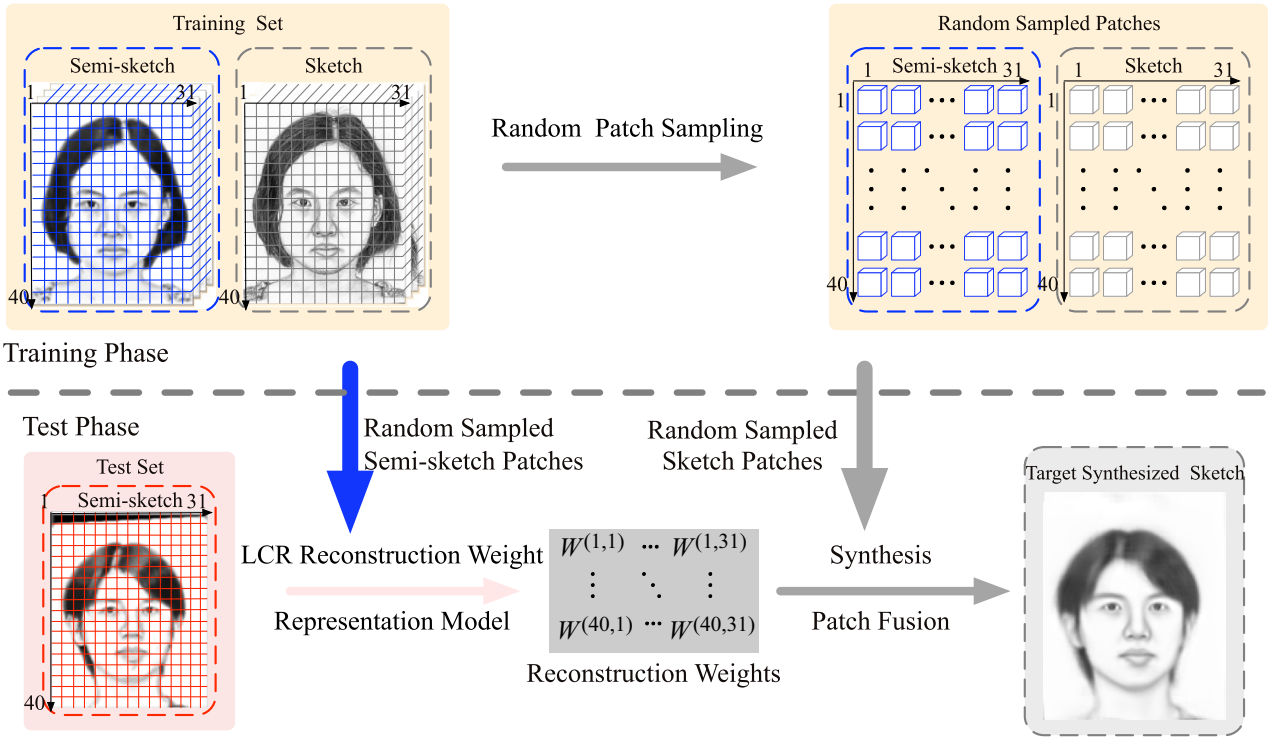


Fig. 7. Illustration of the synthesis process using the RSLCR method [3] and preprocessed data.

dashed line. We use $Rect_g$(green) and $Rect_r$(red) to denote these two rectangles. As we can see in the Fig. 6, $Rect_g$ and $Rect_r$ have covered almost all facial features required to strengthen the supervision, including the smile folds and under-eye bags. The images in the photo–sketch dataset are geometrically aligned relying on two eye centers. In addition, all images in the dataset are of the size 250×200. Therefore, if $Rect_g$ and $Rect_r$ are set to the suitable size, even if the position of these two rectangles is fixed they can cover almost all facial features in any image that is in the dataset. For the datasets we use in this paper, the position and size of $Rect_g$ and $Rect_r$ are set as follows:

$$Rect_g = image(30, 90, 140, 80) \tag{2}$$

$$Rect_r = image(70, 90, 60, 160) \tag{3}$$

where $image(x, y, w, h)$ represents a rectangle area in image, this rectangle's top left corner coordinates are (x, y), the width of this rectangle is w and the height of this rectangle is h.

To preserve more details and generate a delicate semi-sketch, this paper adds two local loss functions: $\mathcal{L}_{Rect_g}$ and $\mathcal{L}_{Rect_r}$.

$$\mathcal{L}_{Rect_g} = \frac{1}{N}\sum_{i=1}^{N} \|\mathcal{X}_i(Rect_g) - \mathcal{S}_i(Rect_g)\|_1 \tag{4}$$

$$\mathcal{L}_{Rect_r} = \frac{1}{N}\sum_{i=1}^{N} \|\mathcal{X}_i(Rect_r) - \mathcal{S}_i(Rect_r)\|_1 \tag{5}$$

Thus, the total loss function of our preprocessing method can be formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{global} + \mathcal{L}_{Rect_g} + \mathcal{L}_{Rect_r} \tag{6}$$

For the dataset which the image is not aligned and the image size is not 250×200, we need to translate, rotate, and scale to align all photos and sketches by the centers of the two eyes. Then we need to average of all the aligned photos in the training set to get an average photo. The approximate position of the rectangular area is selected on the average photo based on the position of the eyes. As mentioned above, the key of rectangular area selection is to cover important facial features. The size of the rectangular area is important for the supervision of facial features. The effect of the size of the rectangular area on the generation of semi-sketches will be detailed in Section 4.3.

3.2. Sketch synthesis via exemplar-based method

After the preprocessing stage, we obtain the training semi-sketches and test semi-sketches with similar distributions. But these semi-sketches still have some noise and lack texture details. The main task at the sketch synthesis stage is to erase the noise and add more texture details. At this stage, we directly employ the existing exemplar-based methods to synthesize target sketches from semi-sketches, since these methods have achieved impressive performance in synthesizing target sketches from photos. Neighbor selection and reconstruction weight computation are the two main parts in the exemplar-based method. Fig. 7 shows the process of synthesizing a sketch using the RSLCR method [3] and the preprocessed data. The RSLCR method is an exemplar-based method which aims to speed up the synthesis while maintaining the high quality of the synthesized results. As shown in Fig. 7, the RSLCR method applies an offline random sample strategy instead of the online searching for neighbors in the training phase, which greatly improves the synthesis speed. In the test phase, the locality constraint is imposed on the reconstruction weight representation, which improves the synthesis quality.

In the following, we take the RSLCR method as an example of the exemplar-based method in the sketch synthesis stage to describe our method. In the preprocessing stage, the training photos and the test photos are converted to the training semi-sketches and test semi-sketches by the pFCN. In the synthesis stage, the RSLCR method is used to synthesis sketches. In the training phase, the *semi-sketch*-sketch pairs in training dataset are divided into $N = 40 \times 31$ patches with even size. At each patch position (i, j),

$n_{rs} = 800$ pairs of training semi-sketch patches and training sketch patches are randomly sampled, $i \in \{1, 2, \ldots, 40\}$, $j \in \{1, 2, \ldots, 31\}$. Then N PCA projection matrices $E(i, j)$ are computed to reduce the dimension of the training semi-sketch patches. The dimension reduced training semi-sketch patches at the patch position (i, j) are denoted as $X^{(i,j)}$. Their corresponding training sketch patches are denoted as $Y^{(i,j)}$. In the test phase, the test semi-sketch is divided into N patches according to the same way as the training *semi-sketch*-sketch pairs have been divided. The N PCA projection matrices $E(i, j)$ computed in the training phase are used to reduce the dimension of the testing semi-sketch patches. The dimension reduced test semi-sketch patch at the patch position (i, j) is denoted as $x^{(i,j)}$. Its corresponding target sketch patch is denoted as $y^{(i,j)}$. At each patch position (i, j), $X^{(i,j)}$ and $x^{(i,j)}$ are fed to the locality constraint (LCR) based reconstruction weight representation model to get the weight representation $W^{(i,j)}$. Then the target sketch patch at position (i, j) can be synthesized: $y^{(i,j)} = Y^{(i,j)}W^{(i,j)}$. The whole target sketch can obtain by arranging all target patches. The above parameters are all from [3].

4. Experiment and analysis

In this section, we first introduce the datasets and the evaluation criteria in Section 4.1. In Section 4.2, we describe the necessary implementation details. Section 4.3 discusses the setting of loss function for pFCN. Afterwards, we employ the well-trained pFCN in the preprocessing stage and adopt four state-of-the-art exemplar-based methods in the synthesis stage to synthesize the target sketches. Section 4.4 provides a qualitative comparison with the state-of-the-art methods. Objective image quality assessment and sketch-based face recognition experiments are carried out as detailed in Sections 4.5 and 4.6 to demonstrate the effectiveness of the proposed two-stage method. Subsequently, the experiments on cross-dataset are conducted to demonstrate the effectiveness of the proposed pFCN preprocessing in enhancing the generalization ability of the exemplar-based method (see Section 4.7).

4.1. Dataset and evaluation criteria

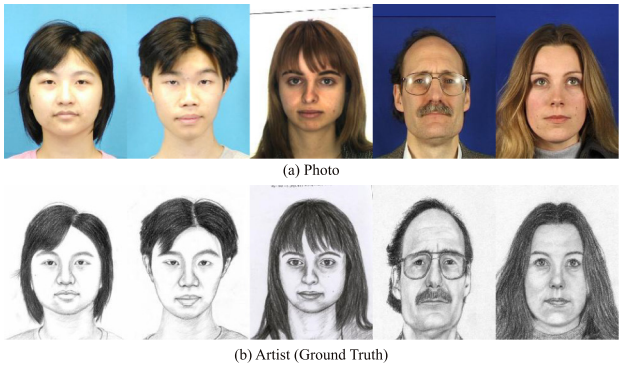
4.1.1. Dataset

To demonstrate the effectiveness of our proposed method, we carried out the experiments on the CUHK Face Sketch dataset (CUFS) [2], which is widely used in face sketch synthesis and recognition [9,35]. This dataset consists of three sub-databases (CUHK student dataset, AR dataset, and XM2VTS dataset) with a total of 606 samples. For each sample, there is a sketch drawn by an artist based on a photo taken in a frontal pose, under normal lighting conditions. Of the 606 samples, 188 are from the Chinese University of Hong Kong (CUHK) student dataset, 123 samples are from the AR dataset [36] and 295 samples are from the XM2VTS dataset [37]. Samples in the XM2VTS dataset are different in age, skin and hairstyles. Some of the photo–sketch pairs from the three sub-databases are shown in Fig. 8.

The settings for the training set and test set are the same as in [3]. Of the 188 samples in the CUHK student dataset, 88 are selected for training, and the remaining 100 samples are taken as the test set. For AR dataset, we take 80 samples as training samples and the remaining 43 samples are used as the test set. From the XM2VTS dataset, 100 samples are chosen for training and the remaining 195 for testing.

4.1.2. Evaluation criteria

The structure similarity index metric (SSIM) [38] is adopted as evaluation criteria in this paper to objectively evaluate the perceptual quality of the synthesized sketches. In recent years, SSIM has become the prevalent metric in sketch face synthesis. The quality



(a) Photo

(b) Artist (Ground Truth)

Fig. 8. Example face photo–sketch pairs in the CUFS dataset. The first and second columns are from CUHK student dataset, the third column is from AR dataset, and the fourth and fifth columns are from the XM2VTS dataset.

of a synthesized sketch can be assessed by computing the SSIM index between itself $\hat{\mathcal{S}}$ and its corresponding hand-drawn sketch (ground truth) $\mathcal{S}$. The SSIM index between two images can be computed as follows:

$$SSIM(\hat{\mathcal{S}}, \mathcal{S}) = [l(\hat{\mathcal{S}}, \mathcal{S})]^{\alpha}[c(\hat{\mathcal{S}}, \mathcal{S})]^{\beta}[s(\hat{\mathcal{S}}, \mathcal{S})]^{\gamma} \tag{7}$$

where $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ are used to adjust the relative importance of the three components. In this paper, $\alpha = \beta = \gamma = 1$. The three components are computed as follows:

$$l(\hat{\mathcal{S}}, \mathcal{S}) = \frac{2\mu_{\hat{\mathcal{S}}}\mu_{\mathcal{S}} + C_1}{\mu_{\hat{\mathcal{S}}}^2 + \mu_{\mathcal{S}}^2 + C_1} \tag{8}$$

$$c(\hat{\mathcal{S}}, \mathcal{S}) = \frac{2\sigma_{\hat{\mathcal{S}}}\sigma_{\mathcal{S}} + C_2}{\sigma_{\hat{\mathcal{S}}}^2 + \sigma_{\mathcal{S}}^2 + C_2} \tag{9}$$

$$s(\hat{\mathcal{S}}, \mathcal{S}) = \frac{\sigma_{\hat{\mathcal{S}}\mathcal{S}} + C_3}{\sigma_{\hat{\mathcal{S}}}\sigma_{\mathcal{S}} + C_3} \tag{10}$$

where $l(\hat{\mathcal{S}}, \mathcal{S})$ is the luminance comparison function, $c(\hat{\mathcal{S}}, \mathcal{S})$ is the contrast comparison function, and $s(\hat{\mathcal{S}}, \mathcal{S})$ is the structure comparison function. C_1, C_2 and C_3 are three constants to avoid instability when the denominator is very close to zero. $C_i = (K_i L)^2$, $K_i \ll 1$, $i = 1, 2, 3$, L denotes the dynamic range of the pixel values (255 for 8-bit grayscale images). We set $C_3 = \frac{C_2}{2}$, $K_1 = 0.01$, $K_2 = 0.03$. μ represents the mean of the image and σ represents the variance of the image or covariance of two different images.

4.2. Implementation details

We employ PyTorch, a popular deep learning platform, for implementation. Photos in the training set are used directly as inputs and their ground truth as labels to train the proposed pFCN. Note that the two eye centers of all input and label images should be in the same positions, and all input and label images are cropped to the size of 250×200. To ensure the output semi-sketch and the input photo have the same size, we apply padding before the convolutional operation, except when the kernel size is 1×1. For the initialization of our network, the filter weights are filled with values $\omega \sim N(0, 0.01)$, and the biases are filled with zero. We set the learning rate to 0.001, and use Adam optimization algorithm [39] to optimize our model (the weight decay is set to 0.0002). The model is trained on an NVIDIA Titan Xp GPU with 12G memory in 120 epochs.

We feed all the photos in the dataset (both training photos and test photos) to the well-trained pFCN to obtain the training semi-sketches and test semi-sketches.

Fig. 9. Comparison of different loss function strategies. (a) are the synthesized sketches of the FCN method [11], it uses an MSE based loss function to train its model; (b) are from the MSE based total loss function trained pFCN; (c) are from the global loss function trained pFCN; (d) are from the L1 based total loss function trained pFCN.

Table 1
Average SSIM values (%) on CUHK student dataset.

Method	FCN	MSE-pFCN	global-pFCN	L1-pFCN
SSIM(%)	60.94	53.58	56.47	**61.78**

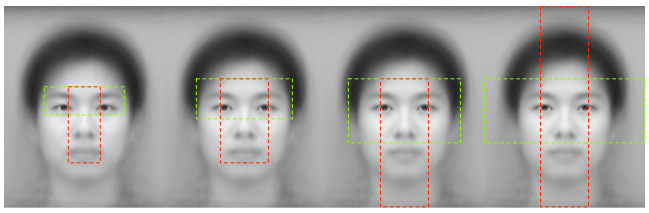


Fig. 10. Illustration of the position and size of the different rectangular areas in the local loss function.

Table 2
Position and size of different rectangular areas.

Rect	x	y	w	h
$Rect1_g$	50	100	100	35
$Rect1_r$	80	100	40	95
$Rect2_g$	40	90	120	50
$Rect2_r$	70	90	60	105
$Rect3_g$	30	90	140	80
$Rect3_r$	70	90	60	160
$Rect4_g$	0	90	200	80
$Rect4_r$	70	0	60	250

Table 3
Average SSIM values (%) under different rectangular areas on CUHK student dataset.

Method	Rect1	Rect2	Rect3	Rect4
SSIM(%)	58.99	61.08	**61.78**	60.70

4.3. Discussion on loss function for pFCN

In this subsection, we discuss the setting of loss function for pFCN on the CUHK student dataset. To verify the effectiveness of the loss function described in Section 3.1.2, we carry out experiments on three aspects.

First, to demonstrate that the L1 loss is more appropriate than MSE loss for this preprocessing task, we use L1 based total loss function (described in Section 3.1.2 formulation (6)) and MSE based total loss function to train pFCN. Replacing the computation of L1-norm in Section 3.1.2 with L2-norm can help to obtain the formulation of MSE based total loss function. The same optimizer and learning rate are used to iterate 120 epochs. The two trained pFCN models are represented as L1-pFCN and MSE-pFCN. Fig. 9(b) reveals that the synthesized results from MSE-pFCN are darker and have more noise in the face area. However, Fig. 9(d) shows that the synthesized results from L1-pFCN highlight the important facial features while weakening the noise. FCN model [11] is trained by MSE based loss function and obtains impressive results; however, it takes several hours to train. The L1-pFCN can synthesize a clearer and more delicate sketch than FCN in several minutes, as shown in Fig. 9(a) and (d), and the synthesized sketches of FCN method [11] are from the results released by Wang et al. [3].

Second, to prove that the use of local loss functions ($\mathcal{L}_{Rect_g}$ and $\mathcal{L}_{Rect_r}$) can improve the quality of the synthesized semi-sketch, we only use the L1 based global loss function $\mathcal{L}_{global}$ to train pFCN. The trained pFCN is denoted as global-pFCN. The synthesized results from global-pFCN can be seen in Fig. 9(c). Comparing Fig. 9(c) with Fig. 9(d) reveals that the addition of two local loss functions can make the image cleaner and refine the facial features, especially for the eyes area.

In addition, the average SSIM values in Table 1 show the superiority of L1-pFCN in an objective way.

Therefore, the L1 based total loss function is more effective for training pFCN than MSE based total loss function and L1 based global loss function, and the proposed architecture and loss function in Section 3.1 are more appropriate than FCN [11] for the preprocessing task.

Third, we discuss the effect of the size of the rectangular area on the generation of semi-sketches. Four groups of rectangular area are used in the loss function respectively. Details on the rectangular area settings are shown in Fig. 10 and Table 2. The denotations in Table 2 and Fig. 10 are consistent with the Section 3.1.2. Table 3 presents the SSIM score corresponding to different rectangular areas. It can be seen from Table 3 that *Rect*3 achieves the best performance in the four sets of rectangular areas. As we can see from Fig. 10, *Rect*1 only covers the important facial features in the average photo. However, the photos in the dataset are roughly aligned. So the *Rect*1 is too small to obtain a good result. With the expansion of the rectangle, the SSIM score is increasing. However, when the rectangle expands to the edge of the image (see Fig. 10(d) *Rect*4), the supervision of the facial features is weakened and the SSIM score is reduced because the rectangle is too large. In order to make the selected rectangle cover the facial features as many photos as possible, the rectangle selected on the average photo face should be properly enlarged. But the oversized rectangle has an adverse effect on the semi-sketch. Therefore, if the pFCN

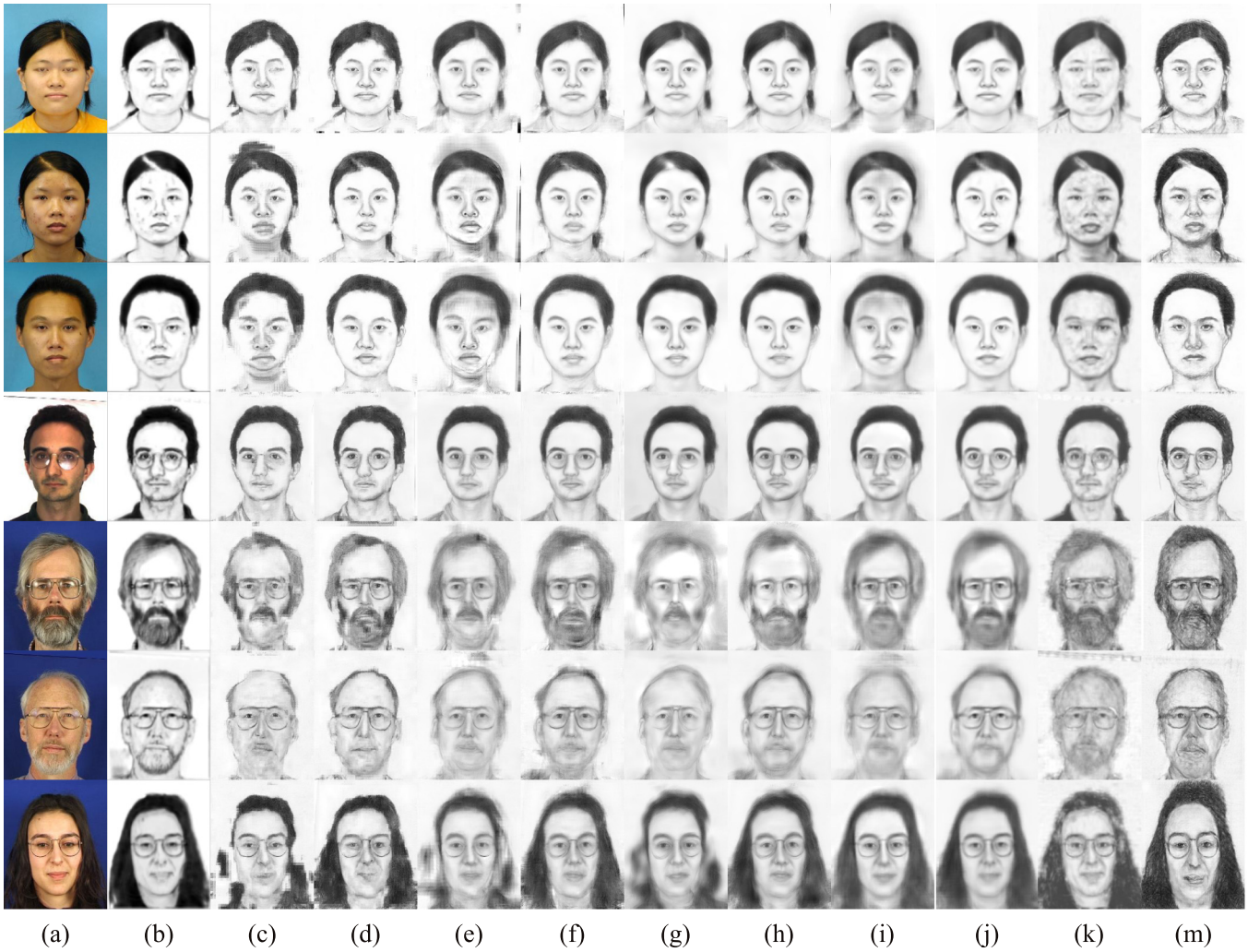


Fig. 11. Comparison of sketches synthesized by different methods. (a): Input Photos; (b): the proposed pFCN; (c): MRF [2]; (d): pFCN+MRF (the proposed two-stage method which uses the MRF method in the sketch synthesis stage); (e): LLE [1]; (f): pFCN+LLE (the proposed two-stage method which uses the LLE method in the sketch synthesis stage); (g): SSD [13]; (h): pFCN+SSD (the proposed two-stage method which uses the SSD method in the sketch synthesis stage); (i): RSLCR [3]; (j): pFCN+RSLCR (the proposed two-stage method which uses the RSLCR method in the sketch synthesis stage); (k): FCN [11]; (m): GAN [40].

needs to be trained in a new dataset, the suitable rectangle can be found by fine-tuning based on the above rules.

In the following experiments, the pFCN is trained by the loss function described in Section 3.1.2. To make the expression more concise, we use pFCN to denote L1-pFCN in the following.

4.4. Photo-to-sketch generation

In this subsection, we evaluate the proposed two-stage approach on the CUFS database. To prove that the two-stage method is more effective than directly using the exemplar-based method or using the pFCN alone, the LLE method [1], MRF method [2], SSD method [13] and RSLCR method [3] as the examples of the exemplar-based methods are applied in the synthesis stage. In addition, we also compare the proposed method with two neural network-based methods: the FCN method [11] and the GAN method [40]. The sketches synthesized using the SSD method and RSLCR method are generated from the source codes provided by the authors. For the MRF method, the synthesized sketches are generated from the codes that are implemented by the author of SSD, while for the LLE method the synthesized sketches are generated from the codes implemented by the author of RSLCR. The sketches generated by the FCN method and GAN method are from the release results by the RSLCR author.

We first use the training photo–sketch pairs to train the pFCN. Then the training photos and test photos are fed to the well-trained pFCN to obtain the training semi-sketches and test semi-sketches (see Fig. 11(b)). In the synthesis stage, we use the *semi-sketch*-sketch pairs and exemplar-based method (LLE, MRF, SSD or RSLCR) to produce the target sketches. The synthesized results of the exemplar-based methods (LLE, MRF, SSD and RSLCR) and the proposed two-stage methods are shown in Fig. 11. As we can see in the results from exemplar-based methods (see Fig. 11(c), (e), (g), (i)), some noise exists in the background and facial skin, some distortions and missing parts in several facial components and some missing parts in the hair area, whereas their corresponding two-stage methods achieve better performance on these aspects (see Fig. 11(d), (f), (h), (j)). The reason behind these improvements can be explained as follows. Most of the exemplar-based methods assume that if two photo patches are similar their corresponding sketch patches also are similar. However, this assumption is stronger for *semi-sketch*-sketch pairs than photo–sketch pairs (see Fig. 2). Therefore, after the preprocessing, exemplar-based methods can select more suitable candidates and calculate more appropriate reconstruction weights to synthesize the test patch. In addition,

Table 4
Average SSIM values (%) on the CUFS database.

Method	pFCN	LLE	LLE[a]	MRF	MRF[a]	SSD	SSD[a]	RSLCR	RSLCR[a]	FCN	GAN	stack-CA-GAN
SSIM(%)	53.85	52.58	**53.26**	51.32	**51.99**	54.20	**55.15**	55.72	**56.10**	52.14	49.39	52.66

[a] Denotes the proposed two-stage method.

for the non-facial components (like the glasses and collar), the two-stage method also performs better than the exemplar-based method because the contours and structure of these components are emphasized in the preprocessing stage (e.g., collar in the first row of Fig. 11 and glasses in the fourth row of Fig. 11).

Some results of neural network-based methods (the proposed pFCN method, FCN method, and GAN method) are shown in Fig. 11(b), (k), (m). Compared to the exemplar-based approaches, the neural network-based methods can generate structurally complete sketches (no missing part). As we can see, the semi-sketch generated by the proposed pFCN method is cleaner than the sketch generated by the FCN method. The GAN method can produce more stylistic results but it also introduces many artifacts. Although the semi-sketch already has some characters of the sketch image, there are still have some shortcoming. In the second row of Fig. 11, some of the acne scars in the photo are retained in the semi-sketch. However, these acne marks are usually ignored by the artist (see Fig. 1(e)). In the last row of Fig. 11, the semi-sketch is more like a photo rather than a sketch. These problems are alleviated after the synthesis stage (see Fig. 11(d), (f), (h), (j)). This proves that the exemplar-based methods in the synthesis stage can convert semi-sketches into the target sketches which have a more similar style with the hand-drawn sketches.

4.5. Objective image quality assessment

The ability to achieve high visual quality results is an important criterion for evaluating a sketch face synthesis method. In this subsection, we utilize the SSIM to evaluate the perceptual quality of synthesized sketches by different methods on the CUFS database. Table 4 shows the average SSIM scores on test dataset in the CUFS database. As we can see, the semi-sketch generated by the proposed pFCN method obtains higher SSIM value than the results of the other three neural network-based methods (the FCN method, the GAN method, and the stack-CA-GAN method [33]). This proves that the architecture and loss function of the proposed pFCN is effective.

Table 4 also shows that the proposed two-stage method achieves higher quality results than its corresponding exemplar-based method. When we use the LLE method and the MRF method in the synthesis stage, the SSIM values of the final results are lower than the semi-sketches generated by the pFCN method. However, this does not mean that the synthesis stage in the two-stage method does not work. Because the two-stage method obtains a better performance on face recognition than using the exemplar-based method or the pFCN method alone, which indicates that the two-stage method can better preserve the identification information (see more details in Section 4.6).

4.6. Sketch-based face recognition

Due to its significant applications in assisting law enforcement, the performance on sketch-based face recognition is also widely used to evaluate the quality of a synthesized sketch. In the sketch-based face recognition task, the hand-drawn sketch is taken as the probe to search the most similar synthesized sketch in the gallery. In this subsection, we apply the null-space linear discriminant analysis (NLDA) [41] to conduct the face recognition experiments. Following the settings in [3], from the 388 samples in the CUFS database, we randomly select 150 synthesized sketches and their ground-truth sketches (hand-drawn by artists) as a training set to train the classifier. The remaining 238 sketches are used as a test set for recognition accuracy statistics. We repeat the face recognition experiment 20 times by randomly splitting the data.

In this subsection, eight state-of-the-art methods are compared. The synthesized sketches of MWF method [12] are from the release results by the RSLCR [3] author. Since the model of the stack-CA-GAN is not available, we only compared the result published in the paper. Table 5 represents the face recognition rates with different reduced dimensions by NLDA. As we can see in Table 5, although the pFCN method achieves higher recognition rate than some of the exemplar-based methods, the two-stage method always achieves higher results than both pFCN and its corresponding exemplar-based method. Therefore, the combination of pFCN and exemplar-based method can facilitate and improve the results from each. Moreover, MWF method introduces the linear combination into the MRF model, which greatly improves the quality of the synthesized results. However, with the help of pFCN, MRF method can perform better than MWF. In addition, the sketches synthesized by the LLE method and the SSD method have lower recognition accuracy than the sketches generated by four neural network-based methods (pFCN, FCN, GAN, and stack-CA-GAN). But their corresponding two-stage methods have better performance on face recognition than these neural network-based methods.

Fig. 12 gives detailed variations of the recognition rate against variations of the reduced number of dimensions by NLDA. It also objectively demonstrates the superiority of the proposed two-stage method compared to the corresponding exemplar-based method or the pFCN method.

4.7. Experiments on the cross-dataset

To verify that the proposed two-stage method has stronger generalization ability than its corresponding exemplar-based method, we take the RSLCR method as the example of the exemplar-based method to conduct the experiment on cross-dataset. The reason for using the RSLCR method here is that its performance is the best among the above-mentioned exemplar-based methods and sketch synthesis is fast.

In this subsection, we take 88 samples from the CUHK student dataset as the training set and 195 samples from the XM2VTS dataset as the test set. Some of the synthesized results are shown in Fig. 13. As we can see from Fig. 13(b), the RSCLR method still performs well in the facial area, especially for the eyes, nose and mouth. However, for the non-facial area (e.g., hair and glasses), the synthesized results by the RSLCR method are unsatisfactory. Fig. 13(c) shows that the two-stage method achieves a cleaner structure and handles the non-facial area better. This is because the pFCN preprocessing converts the different distributions of training samples and test samples into similar distributions, which mitigates the impact of the cross-dataset on the exemplar-based method.

To assess the quality of the synthesized results objectively, the SSIM scores of the synthesized sketches are computed. Table 6 shows the average SSIM scores of the two methods on cross-dataset. Fig. 14 gives the statistics of SSIM scores on cross-dataset. The horizontal axis indicates the SSIM score from 0 to 1. The vertical axis shows the percentage of synthesized sketches whose SSIM

Table 5
The face recognition accuracy (%) of the compared methods with different reduced dimensions by NLDA.

Dim	pFCN	LLE	LLE[a]	MRF	MWF	MRF[a]	SSD	SSD[a]	RSLCR	RSLCR[a]	FCN	GAN	stack-CA-GAN
5	57.10	48.09	**71.89**	44.71	50.08	**67.42**	46.94	**72.50**	70.11	**72.63**	66.33	63.30	–
10	81.54	70.48	**88.27**	66.81	73.22	**85.61**	70.32	**88.91**	87.58	**89.49**	86.41	82.10	–
20	90.85	80.96	**93.75**	76.86	82.87	**91.33**	79.81	**94.49**	93.35	**94.92**	92.42	88.96	–
50	94.55	88.16	**96.62**	84.39	89.87	**95.05**	87.93	**97.31**	96.97	**97.66**	95.85	92.26	–
100	95.48	90.45	**97.37**	86.86	92.50	**95.90**	90.11	**97.90**	97.71	**98.30**	96.78	93.06	–
149	95.80	91.20	**97.39**	87.55	93.01	**96.25**	90.56	**98.01**	98.14	**98.38**	97.10	93.30	95.64

[a] Denotes the proposed two-stage method.

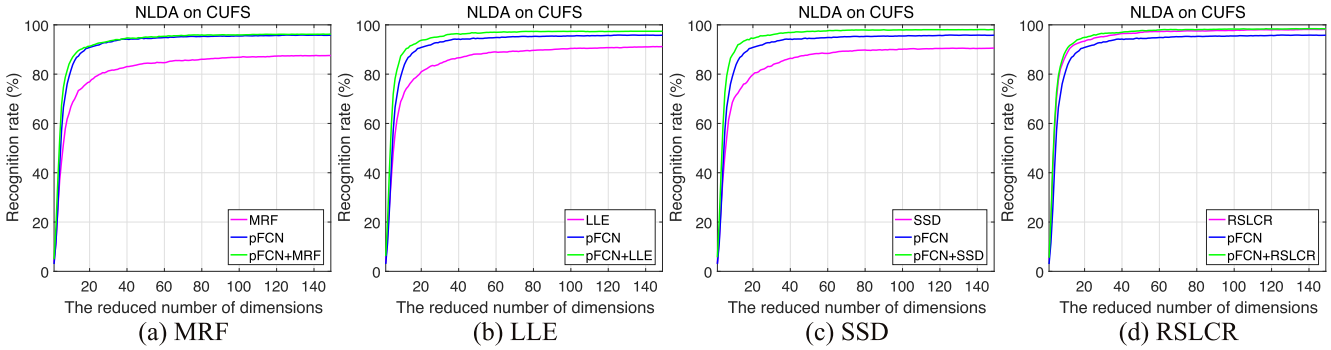


Fig. 12. Variations of the recognition rate against variations of the reduced number of dimensions by NLDA on the CUFS dataset. (a) the comparison of the MRF method [2] and its corresponding two-stage method; (b) the comparison of the LLE method [1] and its corresponding two-stage method; (c) the comparison of the SSD method [13] and its corresponding two-stage method; (d) the comparison of the RSLCR method [3] and its corresponding two-stage method.

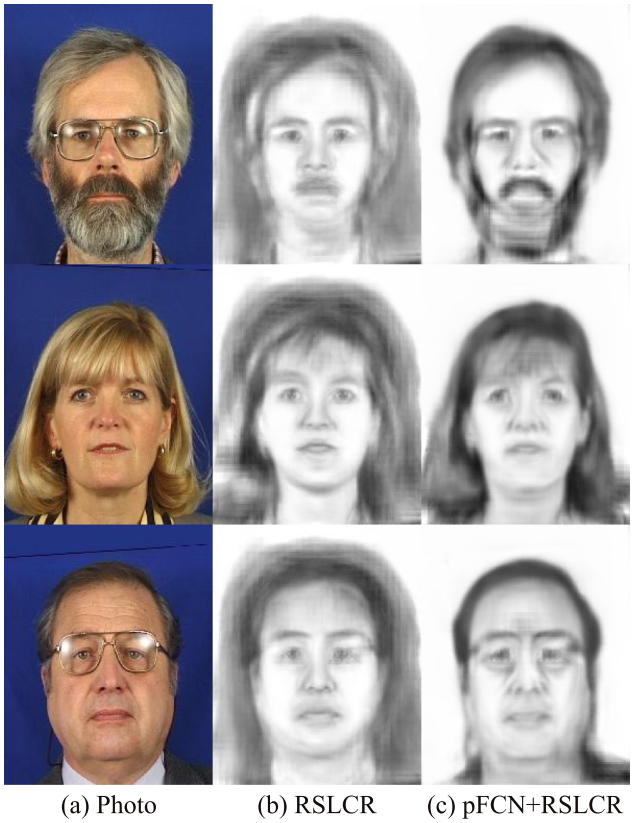


Fig. 13. Synthesized sketches on the cross-dataset by RSLCR [3] and the proposed two-stage method.

Table 6
Average SSIM values (%) on the cross-dataset.

Method	RSLCR	pFCN+RSLCR
SSIM(%)	38.11	**48.04**

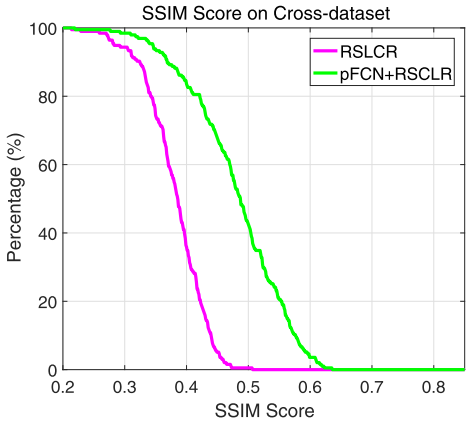


Fig. 14. Statistics of SSIM values (%) on the cross-dataset.

scores are not smaller than the score marked on the horizontal axis. Both Table 6 and Fig. 14 show that the proposed two-stage method produces higher quality results on cross-dataset than the RSLCR method.

In addition, we conduct a sketch-based face recognition experiment to further demonstrate the superiority in the generalization ability of the two-stage method. As in Section 4.6, NLDA is applied to carry out the face recognition experiment. We randomly split the 195 synthesized sketches into a training set (90 synthesized sketches and their ground-truth) and a test set (105 synthesized sketches). We repeat each face recognition experiment 20 times by randomly splitting the data. Fig. 15 presents the variations of the recognition rate against variations of the reduced number of dimensions. The face recognition accuracy also indicates that the proposed two-stage method is superior to the RSLCR method on the cross-dataset.

Therefore, pFCN preprocessing can improve the generalization ability of the exemplar-based method.

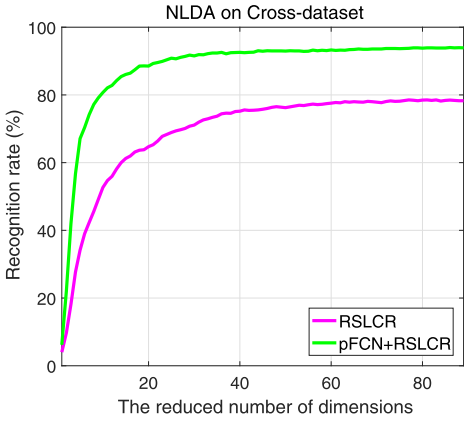


Fig. 15. Variations of the recognition rate against variations of the reduced number of dimensions by NLDA on the cross-dataset.

5. Conclusion

In this paper, we proposed a simple but effective method for sketch face synthesis, which aims to improve the quality of synthesized sketches and enhance the generalization ability of exemplar-based methods. The proposed approach is composed of two stages (preprocessing stage and synthesis stage). In the preprocessing stage, an eight-layer fully convolutional neural network (pFCN) converts the photos to the semi-sketches. In the synthesis stage, an exemplar-based method is employed to convert the semi-sketches to target sketches. Specifically, we design a simple loss function to train the pFCN and generate impressive semi-sketches. Multiple experiments based on four state-of-the-art exemplar-based methods (MRF, LLE, SSD and RSLCR) demonstrate the effectiveness of the proposed two-stage method. In addition, from the experiments on cross-dataset, we find that using the proposed pFCN as preprocessing can improve the generalization ability of the exemplar-based method.

Declaration of Competing Interest

None.

Acknowledgments

This work has been supported by National Natural Science Foundation of China (61203261; 61876099), Shenzhen Science and Technology Research (JCY20170307093018753; JCYJ20180305164401921), Foundation of Ministry of Education Key Laboratory of System Control and Information Processing (Scip201801), Foundation of Key Laboratory of Intelligent Computing & Information Processing of Ministry of Education (2018ICIP03), Foundation of State Key Laboratory of Integrated Services Networks (ISN20-06) and The Fundamental Research Funds of Shandong University (2018JCG07). We would also like to thank the CUHK Face Sketch Database from Multimedia Laboratory of The Chinese University of Hong Kong.

References

[1] Q. Liu, X. Tang, H. Jin, H. Lu, S. Ma, A nonlinear approach for face sketch synthesis and recognition, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, 2005, pp. 1005–1010, doi:10.1109/CVPR.2005.39.
[2] X. Wang, X. Tang, Face photo-sketch synthesis and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 1955–1967, doi:10.1109/TPAMI.2008.222.
[3] N. Wang, X. Gao, J. Li, Random sampling for fast face sketch synthesis, Pattern Recognit. 76 (2018) 215–227, doi:10.1016/j.patcog.2017.11.008.
[4] S. Ouyang, T. Hospedales, Y.-Z. Song, X. Li, C.C. Loy, X. Wang, A survey on heterogeneous face recognition: sketch, infra-red, 3d and low-resolution, Image Vis. Comput. 56 (2016) 28–48.
[5] N. Wang, J. Li, D. Tao, X. Li, X. Gao, Heterogeneous image transformation, Pattern Recognit. Lett. 34 (1) (2013) 77–84, doi:10.1016/j.patrec.2012.04.005.
[6] M. Zhu, N. Wang, A simple and fast method for face sketch synthesis, in: Proceedings of the International Conference on Internet Multimedia Computing and Service, 2016, pp. 168–171.
[7] X. Tang, X. Wang, Face sketch synthesis and recognition, in: IEEE International Conference on Computer Vision, 1, 2003, pp. 687–694, doi:10.1109/ICCV.2003.1238414.
[8] N. Wang, W. Zha, J. Li, X. Gao, Back projection: an effective postprocessing method for gan-based face sketch synthesis, Pattern Recognit. Lett. 107 (2018) 59–65, doi:10.1016/j.patrec.2017.06.012.
[9] M. Zhu, N. Wang, X. Gao, J. Li, Deep graphical feature learning for face sketch synthesis, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2017, pp. 3574–3580, doi:10.24963/ijcai.2017/500.
[10] D. Zhang, L. Lin, T. Chen, X. Wu, W. Tan, E. Izquierdo, Content-adaptive sketch portrait generation by decompositional representation learning, IEEE Trans. Image Process. 26 (1) (2017) 328–339, doi:10.1109/TIP.2016.2623485.
[11] L. Zhang, L. Lin, X. Wu, S. Ding, L. Zhang, End-to-end photo-sketch generation via fully convolutional representation learning, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, 2015, pp. 627–634.
[12] H. Zhou, Z. Kuang, K.K. Wong, Markov weight fields for face sketch synthesis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1091–1097, doi:10.1109/CVPR.2012.6247788.
[13] Y. Song, L. Bao, Q. Yang, M.-H. Yang, Real-time exemplar-based face sketch synthesis, in: Proceedings of European Conference on Computer Vision, 2014, pp. 800–813.
[14] I. Berger, A. Shamir, M. Mahler, E. Carter, J. Hodgins, Style and abstraction in portrait sketching, ACM Trans. Graph. 32 (4) (2013) 55.
[15] H. Chen, Y.-Q. Xu, H.-Y. Shum, S.-C. Zhu, N.-N. Zheng, Example-based facial sketch generation with non-parametric sampling, in: IEEE International Conference on Computer Vision, 2, 2001, pp. 433–438, doi:10.1109/ICCV.2001.937657.
[16] Z. Xu, H. Chen, S. Zhu, J. Luo, A hierarchical compositional model for face representation and sketching, IEEE Trans. Pattern Anal. Mach. Intell. 30 (6) (2008) 955–969, doi:10.1109/TPAMI.2008.50.
[17] H. Chen, L. Liang, Y.-Q. Xu, H.-Y. Shum, N.-N. Zheng, Example-based automatic portraiture, in: The 5th Asian Conference on Computer Vision, 2002.
[18] X. Tang, X. Wang, Face sketch recognition, IEEE Trans. Circuits Syst. Video Technol. 14 (1) (2004) 50–57, doi:10.1109/TCSVT.2003.818353.
[19] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326, doi:10.1126/science.290.5500.2323.
[20] D. De Ridder, R.P. Duin, Locally Linear Embedding for Classification, Tech. Rep. PH-2002-01, Pattern Recognition Group, Dept. of Imaging Science & Technology, Delft University of Technology, Delft, the Netherlands, 2002.
[21] W. Zhang, X. Wang, X. Tang, Lighting and pose robust face sketch synthesis, in: Proceedings of European Conference on Computer Vision, 2010, pp. 420–433.
[22] Y. Song, J. Zhang, L. Bao, Q. Yang, Fast preprocessing for robust face sketch synthesis, in: International Joint Conference on Artificial Intelligence, 2017, pp. 4530–4536.
[23] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, J. Li, Multiple representations-based face sketch–photo synthesis, IEEE Trans. Neural Netw. Learn. Syst. 27 (11) (2016) 2201–2215, doi:10.1109/TNNLS.2015.2464681.
[24] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149, doi:10.1109/TPAMI.2016.2577031.
[25] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: integrated recognition, localization and detection using convolutional networks, 2014, pp. 1–16.
[26] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587, doi:10.1109/CVPR.2014.81.
[27] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823, doi:10.1109/CVPR.2015.7298682.
[28] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems 25, 2012, pp. 1097–1105.
[29] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 640–651, doi:10.1109/TPAMI.2016.2572683.
[30] F. Liu, C. Shen, G. Lin, Deep convolutional neural fields for depth estimation from a single image, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5162–5170, doi:10.1109/CVPR.2015.7299152.
[31] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2016) 295–307, doi:10.1109/TPAMI.2015.2439281.
[32] C. Chen, X. Tan, K.K. Wong, Face sketch synthesis with style transfer using pyramid column feature, in: IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 485–493, doi:10.1109/WACV.2018.00059.
[33] F. Gao, S. Shi, J. Yu, Q. Huang, Composition-aided sketch-realistic portrait generation, arXiv:1712.00899 (2017).

[34] S. Liu, J. Yang, C. Huang, M. Yang, Multi-objective convolutional learning for face labeling, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3451–3459, doi:10.1109/CVPR.2015.7298967.
[35] C. Peng, X. Gao, N. Wang, J. Li, Superpixel-based face sketch–photo synthesis, IEEE Trans. Circuits Syst. Video Technol. 27 (2) (2017) 288–299, doi:10.1109/TCSVT.2015.2502861.
[36] A. Martínez, R. Benavente, The AR Face Database, Technical Report, Computer Vision Center, 1998.
[37] K. Messer, J. Matas, J. Kittler, K. Jonsson, Xm2vtsdb: the extended m2vts database, in: Second International Conference on Audio and Video-based Biometric Person Authentication, 1999, pp. 72–77.
[38] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612, doi:10.1109/TIP.2003.819861.
[39] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv:1412.6980 (2014).
[40] P. Isola, J. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5967–5976.
[41] L.-F. Chen, H.-Y.M. Liao, M.-T. Ko, J.-C. Lin, G.-J. Yu, A new LDA-based face recognition system which can solve the small sample size problem, Pattern Recognit. 33 (10) (2000) 1713–1726, doi:10.1016/S0031-3203(99)00139-9.

Dan Lu was born in Shandong, China, in 1994. She received the B.S. degree in School of Control Science and Engineering, Shandong University, Jinan, China, in 2016. She is pursing the M.S. degree in Control Science and Engineering at the School of Control Science and Engineering, Shandong University, Jinan, China. Her current research interests include machine learning, deep learning and sketch face recognition.

Zhenxue Chen was born in Shandong, China, in 1977. He received the B.S. degree in automatic from School of Electrical Engineering and Automation at Shandong Institute of Light Industry, Jinan, China, in 2000, the M.S. degree in computer science from School of Information Science and Engineering at Wuhan University of Science and Technology, Wuhan, China, in 2003, and the Ph.D. degree in pattern recognition and intelligent systems from Institute of Image Recognition and Artificial Intelligence at Huazhong University of Science and Technology, Wuhan, China, in 2007. Since 2007, Dr. Chen has been an associate professor with the School of Control Science and Engineering, Shandong University. Now, his main areas of interest include image processing, pattern recognition, and computer vision, with applications to face recognition. Dr. Chen has published over 100 papers in refereed international leading journals/conferences such as IEEE Trans. on Industrial Informatics, IEEE Trans. on Circuits and Systems for Video Technology, IEEE Trans. on Intelligent Transportation Systems, Neurocomputing, Neural Computing and Applications and SP-IC.

Q. M. Jonathan Wu (M'92–SM'09) 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 papers 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. Dr. Wu holds the Tier 1 Canada Research Chair in Automotive Sensors and Information Systems. He was Associate Editor for IEEE Transactions on Systems, Man, and Cybernetics Part A, and the International Journal of Robotics and Automation. Currently, he is an Associate Editor for the IEEE Transaction on Neural Networks and Learning Systems and the journal of Cognitive Computation. He has served on technical program committees and international advisory committees for many prestigious conferences.

Xuetao Zhang was born in Shandong, China, in 1995. He received the B.S. degree in School of Mechanical, Electrical & Information Engineering from Shandong University at Weihai, Weihai, China, in 2016. He is pursing the M.S. degree in Control Science and Engineering at the School of Control Science and Engineering, Shandong University, Jinan, China. His current research interests include machine learning, deep learning and semantic segmentation.