Soft Semantic Representation for Cross-Domain Face Recognition

Chunlei Peng, *Member, IEEE*, Nannan Wang, *Member, IEEE*, Jie Li, and Xinbo Gao, *Senior Member, IEEE*

***Abstract*—The problem of cross-domain face recognition aims to identify facial images obtained across different domains, which attracts increasing attentions because of its wide applications on law-enforcement identification and camera surveillance. The problem is challenging due to the huge domain discrepancy. Despite great progress achieved in recent years, existing algorithms usually fail to fully exploit the semantic information for identifying cross-domain faces, which could be a strong clue for recognition. In this article, we propose an effective algorithm for cross-domain face recognition by exploiting semantic information integrated with deep convolutional neural networks (CNN). We first introduce a soft face parsing algorithm where the boundaries of facial components are measured as probabilistic values. By taking the original face image as the guidance to improve face parsing result, each pixel may belong partially to the facial component to avoid inaccurate segmentation around component boundaries. We then propose a hierarchical soft semantic representation framework for cross-domain face recognition. Both the soft semantic level and contour level deep features obtained via CNN are computed and combined together, which could fully exploit the identical semantic clue among cross-domain faces. We provide extensive experiments to demonstrate that the proposed soft semantic representation algorithm performs superior against state-of-the-art methods.**

***Index Terms*—Cross-domain face recognition, semantic representation, face parsing.**

Manuscript received October 26, 2019; revised April 13, 2020 and July 7, 2020; accepted July 18, 2020. Date of publication July 30, 2020; date of current version August 12, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103202; in part by the National Natural Science Foundation of China under Grant 61806152, Grant 61922066, Grant 61876142, Grant 61671339, Grant 61772402, Grant U1605252, and Grant 61976166; in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2019JM-289; in part by the Key Research and Development Program of Shaanxi under Grant 2020ZDLGY08-08; in part by the China Post-Doctoral Science Foundation under Grant 2018M631124 and Grant 2019T120880; in part by the National High-Level Talents Special Support Program of China under Grant CS31117200001; in part by the Fundamental Research Funds for the Central Universities under Grant JB191502 and Grant JB190117; in part by the China 111 Project under Grant B16037; in part by the Xidian University Intellifusion Joint Innovation Laboratory of Artificial Intelligence; and in part by the Innovation Fund of Xidian University. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Domingo Mery. *(Corresponding authors: Xinbo Gao; Nannan Wang.)*

Chunlei Peng is with the State Key Laboratory of Integrated Services Networks, School of Cyber Engineering, Xidian University, Xi'an 710071, China (e-mail: clpeng@xidian.edu.cn).

Nannan Wang is with the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: nnwang@xidian.edu.cn).

Jie Li is with the Video and Image Processing System Laboratory, School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: leejie@mail.xidian.edu.cn).

Xinbo Gao is with the State Key Laboratory of Integrated Services Networks, School of Electronic Engineering, Xidian University, Xi'an 710071, China, and also with the Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: xbgao@mail.xidian.edu.cn).

Digital Object Identifier 10.1109/TIFS.2020.3013209

I. Introduction

FACE images can be captured from different domains in various tasks. One popular scenario is to match face photos with sketches generated based on eyewitness descriptions of the suspect in law enforcement application. The forensic artist can be invited to create a hand-drawn sketch based on the verbal description, or software can be applied to generate a composite sketch for identifying the suspect. There is huge domain discrepancy between face photos and sketches since the photos are captured via a real-life environment while both hand-drawn and software-generated sketches are created manually with artifact [1]. Another cross-domain face recognition scenario in the camera surveillance application is to match visual spectra images (VIS) with near-infrared images (NIR), which is because the near-infrared images are robust to varying illumination environment. Considering the VIS-NIR faces are captured by different sensors, there is also huge domain discrepancy between them. Other cross-domain face recognition scenarios, such as matching ID photos with live face photos [2] and RGB-D face recognition [3], are widely applied in public security, too. An illustration of cross-domain faces is shown in Fig. 1.

Cross-domain face recognition, also known as heterogeneous face recognition [4], aims to match face images across different domains. Usually a number of face image pairs from two different domains are taken as the training subset, while the rest face image pairs are utilized for testing. Cross-domain face recognition is a challenging task because of the various domain discrepancy. A number of cross-domain face recognition algorithms have been developed in recent years. Early cross-domain face recognition methods usually design hand-crafted architectures to either transform face images into the same domain with face synthesis techniques [5], or directly matching cross-domain faces with the help of a common subspace [6] or hand-crafted feature descriptors [7]. However, they are usually computationally expensive and the performance of hand-crafted methods is often limited. With the great development of deep convolutional neural networks (CNN) in classification problems, an increasing number of deep learning based cross-domain face recognition methods are proposed recently. For example, an end-to-end deep network was applied for cross-domain face synthesis [8] to eliminate the domain discrepancy. Restricted Boltzmann Machines (RBMs) [9], CNN [10], and generative adversarial networks (GAN) [11] are also applied to deal with cross-domain face recognition. However, most of the existing method suffer similar drawback that the semantic information across face images from different domains was not well explored, which according

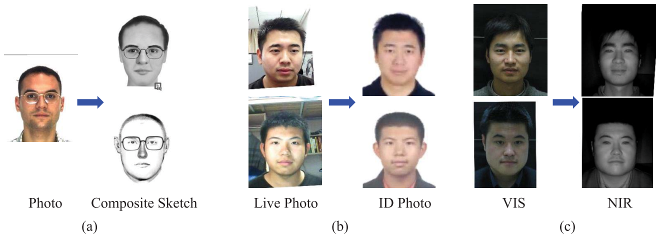


Fig. 1. Example images of cross-domain faces. (a) Photo domain to composite sketch domain; (b) Live photo domain to ID photo domain; (c) VIS to NIR.

to our common sense could be a strong clue for identity recognition. By finding an optimal way to exploit semantic information, the cross-domain face recognition performance can be significantly improved.

Actually, semantic information has been proved to be effective in many fields, such as face labeling [12], face deblurring [13], [14], face hallucination [14], reflection removal [15], face synthesis [16]–[18], and so on. For example, Shen *et al.* [13] exploited semantic cues in face deblurring task with the assumption that face images share key semantic components and the semantic information of a face provides a strong prior for restoration. Zhang *et al.* [17] proposed a generative adversarial network to generate face sketches which embedded background information and semantic information into a deep network based generator and illumination layer respectively. The illumination layer with semantic operation could help cope with illumination variations in photo images. Song *et al.* [14] proposed a facial component guided CNN framework for joint face hallucination and deblurring, where the semantic facial structure is taken into consideration to generate high-resolution faces.

There are several existing techniques which attempted to integrate semantic information into recognition or identification tasks. Kalayeh *et al.* [19] proposed to integrate human semantic parsing in person re-identification task since since a rectangular box may include background and cannot capture deformable human body characteristic. They took Inception-V3 [20] as the backbone architecture for their human parsing model. However, their method was mainly designed for human semantic parsing, while face parsing is a more delicate task. Benefiting from the development of deep learning based facial landmark detection algorithms [21], it can be more efficient to obtain an initial face parsing map under the guidance of these landmarks. Han *et al.* [22] designed a component-based representation approach for matching composite sketches to photos. However, they just cropped the facial components with rectangular boxes roughly and utilized hand-crafted features for matching. The semantic information was not well exploited in [22]. Chen and Ross [23] proposed a semantic-guided generative adversarial network (SG-GAN) to regularize GAN training with semantic priors when synthesizing visible images from thermal images. However, the face parsing results in [23] can be inaccurate around their component boundaries and the recognition performance in [23] heavily depends on the quality of synthesized images. However, existing attempts on applying semantic information suffer the same drawback that a hard classification is performed for each pixel in face images, which leads to hard boundaries of the semantic components. Therefore, the segmentation of the pixels around component boundaries can be inaccurate in the semantic map, which is ignored in existing methods.

In order to better exploit semantic information for cross-domain face recognition, we propose an effective soft semantic representation based framework (SoftSR) in this article. Considering the drawbacks of existing methods integrated with hard pixel classification based semantic maps, we introduce a soft face parsing algorithm where a soft classification is conducted on the pixels around the facial component boundaries. We take the original face image as a guidance to obtain a soft semantic map, where each pixel may belong partially to a semantic facial component. We further devise a hierarchical soft semantic representation framework for cross-domain face recognition. Deep CNN features are extracted from both the soft semantic component level and the facial contour level, which are then combined to improve the performance of cross-domain face identification task. In this way, the identical semantic information across different domains is fully exploited. The contributions of the article are summarized below:

1) We propose a soft face parsing algorithm to exploit the semantic information for cross-domain face recognition. The pixels belong partially to the component as probabilistic values in our soft semantic map, which can cope with inaccurate segmentation around semantic facial boundaries.
2) We develop a hierarchical soft semantic representation framework integrated with deep convolutional neural networks to improve the performance of cross-domain face recognition.
3) We evaluate the proposed algorithm through extensive experiments on multiple cross-domain face dataset and show that our algorithm performs superior against state-of-the-art methods.

II. Related Work

In this section, we provide a literature review on recent works of cross-domain face recognition, which can be roughly grouped as hand-crafted methods before onset of deep learning, and deep learning based methods.

A. Hand-Crafted Methods

Cross-domain face recognition began with an eigen-transformation based method [24] which transformed face photos into the sketch domain, and conventional face recognition algorithm was applied on matching synthesized sketches with sketches in the target gallery. Wang *et al.* [25] later integrated sparse feature selection with support vector regression for cross-domain face transformation. In recent years, a number of Markov networks based methods [5], [26]–[30] were proposed to transform cross-domain faces into a common domain for recognition. There is a comprehensive survey [31] about the synthesis based methods for cross-domain face recognition. However, the performance of these image transformation based method is limited by the quality of the synthesized faces.

Another solution to cross-domain face recognition is to directly matching them with the help of a common subspace or a domain invariant feature descriptor. Lin and Tang [32] proposed a common discriminant feature extraction method for cross-domain face recognition. Kan *et al.* [33] explored the inter-class and intra-class correlations of cross-domain faces for discriminant metric learning. Hou *et al.* [6] intended to cope with cross-domain face recognition by balancing intrapersonal cross-domain distance and interpersonal cross-domain distance. A number of domain invariant hand-crafted descriptors were designed to directly matching cross-domain faces [34]–[38]. Peng *et al.* [39] went through a number of hand-crafted features for matching face photo with both single sketch and multiple stylistic sketches. Jin *et al.* [40] presented a coupled discriminative feature learning method for cross-domain face recognition. A coupled of image filters were learned to exploit discriminative information from faces captured across different domains. The drawbacks of these hand-crafted matching methods are that discriminative information may lose during the projection to a common subspace or designing hand-crafted descriptors.

B. Deep Learning Based Methods

With the great progress of deep learning, the inherent distribution of raw pixels can be explored directly. An end-to-end cross-domain face synthesis method [8] was proposed to generate high quality synthesized faces. The RBMs, CNN, and GAN were also introduced to deal with cross-domain face recognition task [9], [27], [41]. Galea and Farrugia [42], [43] proposed to fit 3D morphable model to face photos and sketches which could synthesize new images for training data augmentation. It was later extended to a very deep CNN with morphed faces and transfer learning [1] for matching software-generated sketches to photos. Pereira *et al.* [44] investigated the performance of directly applying CNN architectures trained with visible photos to cross-domain recognition scenario, and further proposed to learn domain specific feature detectors for different cross-domain face recognition scenarios. Li *et al.* [45] proposed an age factor guided CNN framework for cross-age face recognition which combined an identity discrimination network with an age discrimination network. By adversarial learning both networks, the obtained deep features can be robust to recognizing faces across age variation. An orthogonal dictionary alignment approach [46] was proposed to cope with low-resolution NIR-VIS face recognition, which also designed a re-ranking technique to improve the performance. Deng *et al.* [47] incorporated mutual component analysis into CNN by taking it as a special fully-connected layer, which could extract domain-independent features for recognition. Nasrabadi *et al.* designed attribute-guided algorithms to cope with thermal to visible face recognition [48] and cross-resolution face recognition [49] tasks. Recently Zhu *et al.* [50] investigated matching ID photos with spot photos. The challenges of recognizing photos from ID domain and spot photo domain include the heterogeneity between ID and spot photos, the bisample issue where only two samples are available for each subject, and the large scale classes data. They proposed a progressive model transferring method via a classification-verification-classification pipeline. Deng *et al.* [51] proposed a two-branch Residual networks architecture with residual compensation module and modality discrepancy loss for cross-domain face recognition, and they achieved promising performances on both sketch-photo matching and NIR-VIS matching tasks. Yu *et al.* [52] recently proposed to apply generative adversarial network for face sketch-photo transformation with a feature level loss. He *et al.* [10], [53]–[56] proposed a number of approaches by introducing image generation for matching NIR faces with VIS faces.

However, most of existing cross-domain face recognition methods did not take the semantic clue into consideration. According to our common sense, the semantic information of facial components usually keep consistent across faces from different domains (except for caricature recognition where there is shape exaggeration). Based on this observation, we intend to build the correlations between cross-domain faces and propose a soft semantic representation approach which is robust to image domains. Different from the component-based approach (CBR) in [22] which used a hard component segmentation and hand-crafted features, *we introduce a soft face parsing algorithm and deep CNN features are extracted for soft semantic representation, which could fully exploit the identical semantic information for cross-domain face recognition.*

III. Proposed Approach

In this section, we present our proposed soft semantic representation for cross-domain face recognition. We denote $X = \{X_1, X_1, \cdots, X_M\}$ and $Y = \{Y_1, Y_1, \cdots, Y_M\}$ as a collection of M face images in two different domains. Our method is composed of two phases. Firstly, a soft semantic face parsing algorithm is applied to the images in both domains, and generates a soft semantic parsing map for each face image. Secondly, we design a hierarchical framework to extract deep CNN features from both the soft semantic level and contour level under the guidance of the soft face semantic map. It is a common strategy to utilize a step-to-step method for cross-domain face recognition problem. For example, in [55] the face images are firstly processed by a face completion method and then fed into a CNN for cross-domain face matching. Reference [40] firstly learn a discriminative face representation to obtain a new image pattern, and then the filtered images are projected into one common subspace for recognition. Similarly, we present a two-phase framework to solve cross-domain face recognition. We will describe each phase in details, and without loss of generality, we take face sketch-photo recognition as the example below.

A. Soft Semantic Face Parsing

Inspired by existing works [12]–[14] on applying semantic face parsing to face-related tasks, we would like to take the semantic face parsing map as a cross-domain clue for our recognition mission. We follow existing methods and categorize facial components into four types as eyebrows, eyes, nose, and mouth. For an face image (it could be a photo or a sketch image), we firstly detect facial landmarks using convolutional experts constrained local models [21]. It is feasible to use the dense facial landmarks to guide the semantic parsing map generation. A simple way is to generate the semantic face parsing map by directly connecting neighboring landmarks around each facial component. The pixels within each polygon are assigned to corresponding facial component category. For example, the pixels with the connected landmarks around left eyebrow are classified as the semantic region for left eyebrow. However, this directly connection operation may cause piecewise linear effect because there are some long-distance neighboring landmarks. To solve this phenomenon, we further apply cubic spline interpolation to smooth the boundary of the component. Finally for each face image, we could generate an initial semantic parsing map where the pixels within facial component regions are marked as 1 while the other pixels as 0.

However, the initial semantic parsing map suffers the same drawback as existing works that a hard pixel classification is performed around the boundaries of components. The pixels on the inner side of the boundary is marked as 1 and outside of the boundary as 0. But the boundary of the component could be inaccurate which is ignored in existing techniques. For example, the boundary of the nose bridge may be uncertain while the above hard pixel classification may provide an inaccurate guidance for our following phase. Therefore, in this article we present a soft semantic face parsing algorithm where the pixels around the facial component boundaries are partially belonging to the component category. *Our motivation is that the semantic parsing map need to be smoothed under the guidance of the original face image. For the visually clear boundaries in original face image (eye corners, for example), the boundaries in our soft semantic parsing map should also be clear. But for the visually uncertain boundaries in original face image (nose bridge, for example), the boundaries in the soft semantic parsing map should be smoothed.*

Firstly, we consider a 1D discrete signal $S = \{s_0, s_1, \cdots, s_i, \cdots, s_n\}$ which need to be filtered under the guidance of another signal $P = \{p_0, p_1, \cdots, p_i, \cdots, p_n\}$. A transformed signal $T = \{t_o, t_1, \cdots, t_i, \cdots, t_n\}$ is computed for each element in S where two elements lie on the same side of a strong edge in the guidance signal P are close, while elements lie on different sides of a strong edge in P are far apart. Motivated by [57], we perform a guided edge-preserving filtering to the signal S via a recursive operation as follows:

$$u_i = (1 - \alpha^{\beta})s_i + (\alpha^{\beta})u_{i-1} \tag{1}$$

Here u_i represents the filtered result. $\alpha = \exp(-\sqrt{2}/\sigma_s)$ is a feedback coefficient with a parameter σ_s, and we assume $\alpha \in [0, 1]$. β is the distance between the neighboring elements of the transformed signal T, which is estimated by $\beta = t_i - t_{i-1}$. When β increases, α^{β} tends to be zero and the recursive operation above comes to an end. Therefore, the strong edge in the guidance signal P will make the elements in transformed signal T be far apart, thus β becomes a large value to stop the recursive filtering. The curve in input signal can then be preserved and vice versa. The elements in the transformed signal T are processed under the guidance of P by calculating the sum of the spatial and intensity differences between two elements:

$$t_i = \sum_{j=0}^{i}(1 + \frac{\sigma_s}{\sigma_r}|p_j - p_{j-1}|) \tag{2}$$

Here parameter σ_r is introduced to balance the smoothness of the filter together with σ_s.

Since an image can be regarded as a 2D signal, the 1D edge-preserving recursive filtering introduced above will be performed separately along each dimension of the image iteratively. For example, the 1D recursive operation will be performed horizontally along each image row and then vertically along each image column. The process of 1D filtering operation has a low computational cost and could work on a face image in real time. In practice, three iterations of the recursive operation could obtain satisfactory result for an image.

In our soft semantic face parsing operation, we take each row/column in the initial semantic parsing map as the input signal S, and each row/column in the input face image as the guidance signal P. Therefore, the edges in the input face image could help guide the smoothing of the semantic parsing map. When there is a clear boundary like eye corner in the input face image, β would be a large value and no smoothing is performed. On the other hand, around unclear boundary like nose bridge β would be a small value, and the recursive smoothing filtering will be performed on the semantic parsing map to generate a soft boundary. An illustration about the initial semantic parsing and proposed soft semantic parsing map is shown in Fig. 2.

Formally, given an input photo X_m with its initial semantic parsing map X_m^{ini}, our soft semantic parsing map X_m^s can be generated as follows:

$$X_m^s = SSFP(X_m, X_m^{ini}, \sigma_r, \sigma_s) \tag{3}$$

where $SSFP$ denotes the soft semantic face parsing operation. There are two parameters σ_s and σ_r here. The performance of our method is influenced by the choice of the parameters. σ_s can be regarded as a spatial parameter introduced in equation (1). It controls the feedback coefficient of the recursive edge-preserving filter in (1). σ_r can be viewed as a range parameter which influence the range area of smoothing operation. When σ_s and σ_r increases, the smoothing phenomenon around the facial component boundaries becomes more obvious as shown in Fig. 4 below. Specificity, when the range parameter σ_r becomes relatively large, the parsing map can be extremely smoothed and useful facial component boundaries will be missing. On contrary, the parsing map will not be over smoothed when the spatial parameter σ_s is very large. Therefore, the default parameter setting of σ_r needs to be a proper value, while σ_s can simply be a large value. Since the computing time of the proposed parsing method is very small as will be discussed in the complexity analysis subsection, tuning the two parameters is not quite time consuming. For

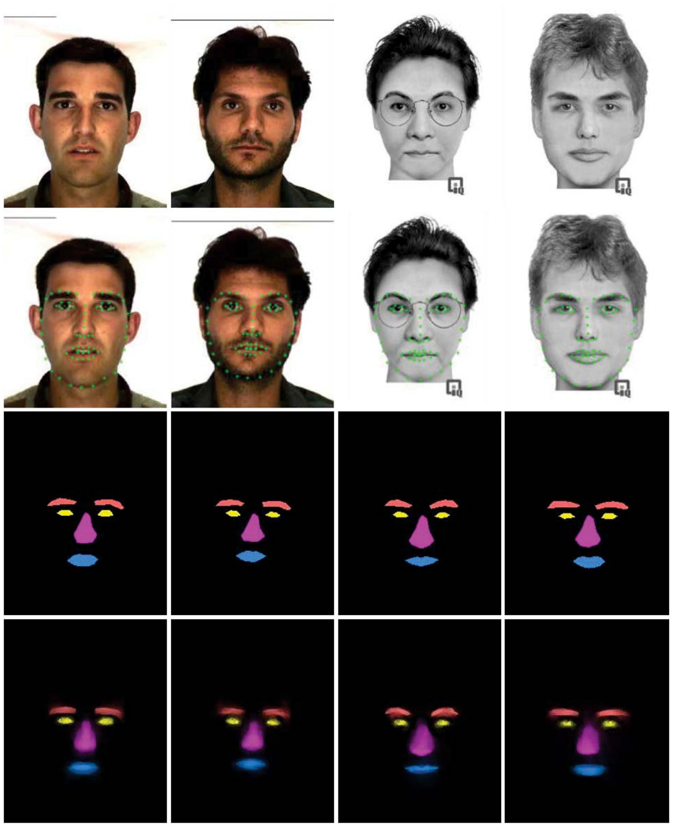


Fig. 2. Examples of semantic face parsing results. There are two photos and two composite sketches shown in the top row. Second row shows the detected facial landmarks in cross-domain face images. Third row demonstrates the initial semantic parsing maps where a hard pixel classification is performed at the boundaries. The last row shows our soft semantic parsing maps where the component boundaries are smoothed under the guidance of the face images in the first row. Here for better demonstration, we mark four facial component categories with different colors.

different cross-domain face recognition tasks, the influence of these parameters is similar. In our experiments satisfactory recognition performance can be obtained for different face images with a default parameter setting. Furthermore, it is normal that the performance of an algorithm is dependent on the parameters, and detailed analysis about the parameters is usually provided such as [10], [40], which will not lead to an uncontrollable situation. We will present detailed discussions about the influence of these two parameters and the way to set them in subsection IV-A, which can be helpful to analyze the influence of these parameters and guide the choice of them for different cross-domain face recognition tasks.

B. Soft Semantic Representation for Cross-Domain Face Recognition

We now present a hierarchical framework based on the soft semantic face parsing to handle the domain discrepancy in cross-domain face recognition. As shown in Fig. 2, our method combines both soft semantic level and contour level deep features. The soft semantic level features could capture comprehensive information from both soft semantic component shapes and component contents, thus are independent of the domain discrepancy. The former can be viewed as a high-level properties of the face by extracting semantic feature at each pixel, while the latter can represent the appearance of each semantic component. We further utilize contour level features from the holistic face to provide a semantic description of the whole face, which are complementary and the fusion of these hierarchical deep features could provide efficient and robust performance on cross-domain face recognition. The soft semantic representation is discussed in detail below.

1) Soft Semantic Level Feature: The usage of high-level semantic features has demonstrated its effectiveness in face labeling [12], face deblurring [13], [14], face synthesis [16]–[18] and many other fields. Considering the semantic information of facial components usually keep consistent across faces from different domains, we extract the soft semantic clue at each pixel position to represent whether the pixel is partially belong to a facial component category or not. We generate a R-dimensional probabilistic vector at each pixel, where R is the number of component categories. This vector represents the probability that the pixel partially belongs to each facial component. We simply concatenate the R dimensional vectors at all pixels together to obtain the soft semantic level shape feature, which is denoted as follows:

$$F_{SS} = f(X_m^s, R) \tag{4}$$

where F_{SS} refers to the soft shape feature which can be easily obtained with the soft semantic parsing map X_m^s as input. In this article, we consider six component categories, including left eyebrow, left eye, right eyebrow, right eye, nose, and mouth. Therefore, R is set to be 6.

In order to represent the appearance of soft facial components, we adopt deep convolutional neural networks on small image patches to reach an image-level decision [58]. We construct an image patch level CNN following the architecture in [59], which is composed of seven convolutional layers, batch normalization layers and ReLU layers. We use the patch descriptor learning dataset [60] to train the network and fine-tune on cross-domain facial patches. We densely sample keypoints within the region of facial components and extract small image patches around these keypoints to generate deep features, which are then concatenated together to form soft semantic level content feature. The progress can be denoted as follows.

$$F_{SC} = f(X_m, X_m^s, \Theta) \tag{5}$$

Here F_{SC} refers to the soft content feature. The soft semantic parsing map X_m^s is applied to locate the region of each soft facial components in the input photo X_m. Θ denotes the learned parameters in the deep network after fine-tuning, which is utilized to extract deep features.

2) Contour Level Feature: Besides the soft semantic level feature which focuses on local components, we can further capture comprehensive information from facial contour, which is robust to the changes of pose and expression. We take the landmarks in initial semantic parsing map X_m^{ini} as the keypoints and extract deep features with the help of the image patch level network introduced above. The extracted deep features are concatenated together as the contour level feature,

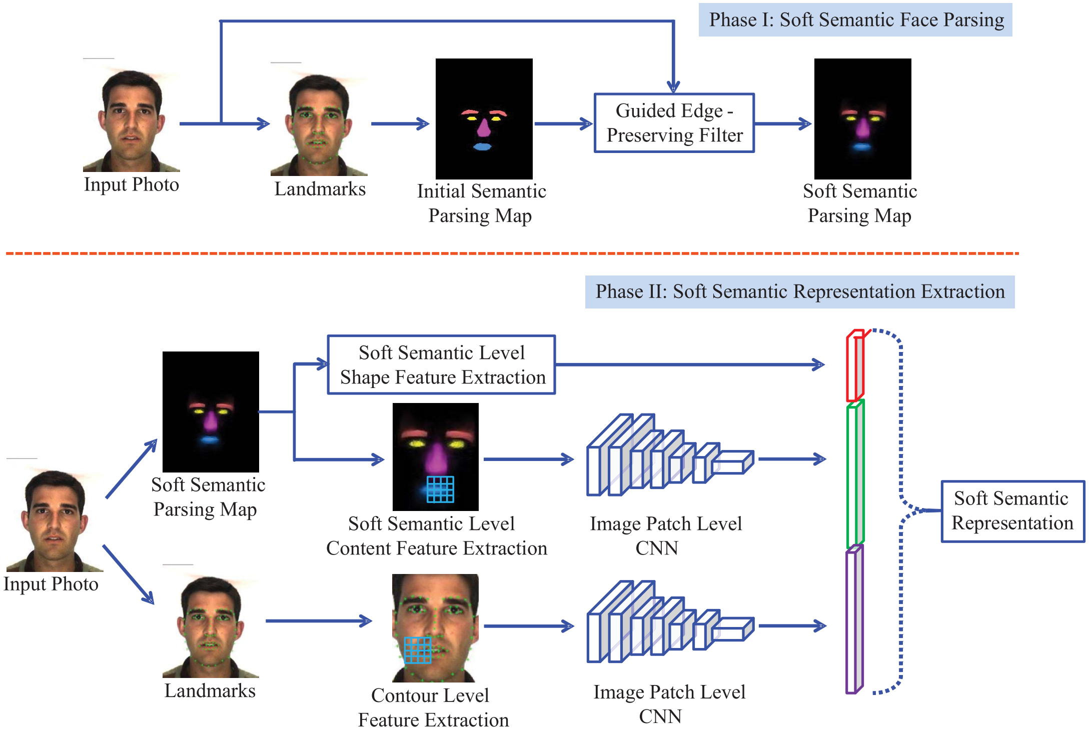

Phase I: Soft Semantic Face Parsing
Input Photo
Landmarks
Initial Semantic Parsing Map
Guided Edge - Preserving Filter
Soft Semantic Parsing Map
Phase II: Soft Semantic Representation Extraction
Soft Semantic Level Shape Feature Extraction
Soft Semantic Parsing Map
Input Photo
Soft Semantic Level Content Feature Extraction
Image Patch Level CNN
Soft Semantic Representation
Landmarks
Contour Level Feature Extraction
Image Patch Level CNN

Fig. 3. Framework of our proposed soft semantic representation. The face images are processed to obtain the soft semantic parsing map in the first phase, and the soft semantic level features and contour level features are extracted respectively. The proposed soft semantic representation is the set of these feature vectors.

which can be formulized as below.

$$F_{HC} = f(X_m, X_m^{ini}, \Theta) \tag{6}$$

where F_{HC} represents the initial hard semantic level contour feature. The contour level feature can provide complementary information to soft semantic level feature, which could fully exploit the identical semantic clue between cross-domain faces.

3) Soft Semantic Representation Fusion: The soft semantic level features F_{SS} and F_{SC}, and the contour level feature F_{HC} are first extracted from each facial component categories. For each semantic facial component, we concatenate the soft semantic level shape feature with the content feature extracted at the key points within this component region together to form one soft semantic level representation $[F_{SS}; F_{SC}]$. Considering there are six component categories, including left eyebrow, left eye, right eyebrow, right eye, nose, and mouth, we can obtain six soft semantic level representations. Furthermore, the deep features extracted at the contour landmarks are concatenated to form the contour level representation. Elements of the six soft semantic level representations and the contour level representation are normalized by the square root, thus the set of these representations form our soft semantic representation. The scheme of our soft semantic representation is shown in Fig. 3.

The cross-domain face recognition is conducted using the set of soft semantic representations. Firstly, we build one classifier for each soft semantic level representation and the contour level representation. Secondly, the scores of these classifiers are fused by simple averaging operation. Since the dimension of F_{SC} and F_{HC} are usually very high, we map them into a discriminant subspace with principal component analysis algorithm and Fisherface [61] algorithm. More advanced classifiers and fusion techniques can be applied here to further improve the performance, which could be investigated in the future. The implementation details of our framework will be introduced in experimental section below.

C. Complexity Analysis

Here we mainly discuss the complexity of the proposed soft semantic face parsing in phase I. Our experiments are conducted on Ubuntu 16.04 with a NVIDIA RTX 2080Ti GPU and 16-GB memory. The computing time of the proposed method is very fast, and it takes only 0.74s for processing one face image. The complexity of the soft semantic face parsing phase is $O(N)$, where N is the number of pixels in the image. Therefore, the proposed soft semantic face parsing phase can be implemented in real-time which means it can be easily applied into other related tasks such as face deblurring, face denoising and face synthesis in the future.

IV. Experiments and Results

In this section, we describe the experimental performance of our proposed method on multiple cross-domain face recognition databases along with the comparison to state-of-the-art methods. Experiments are conducted on publicly available cross-domain face databases, including the photo domain with sketch domain (*e.g.*, e-PRIP [41], PRIP-VSGC [62], UoM-SGFS [1] and a collected forensic sketch database [7]), the live photo domain with the ID photo domain (NJU-ID [63]), and the VIS domain with the NIR domain (CASIA NIR-VIS 2.0 [64]).

The e-PRIP database contains 123 photo-sketch pairs, where the photos are collected from the AR database [65] and the sketches are generated by a composite sketch software. The database is randomly split into two subsets, with 48 pairs used for classifier training and the rest for testing.

The PRIP-VSGC database contains the same 123 photos with the e-PRIP database, while the sketches are created with another different composite sketch software. The same protocol used in e-PRIP database is applied here for evaluation.

The UoM-SGFS database is composed of 600 subjects of which the photos are collected from the FERET database [66]. For each face photo, one sketch is created by a composite sketch software EFIT-V and another sketch is generated by further editing the formal sketch with a software Corel PaintShop Pro-X7. Therefore, the initial composite sketches created by EFIT-V are collected as subset A while the further edited sketches are collected as subset B. The standard evaluation protocol defined in [59] is followed in this article.

The forensic sketch database is composed of 168 forensic sketch-photo pairs collected from real-world law enforcement scenarios. This is a challenging face sketch-photo recognition database since the mug shot photos are captured under various environments while the forensic sketches are drawn according to eyewitness description. We follow the split protocols in [7], [38] where 106 image pairs are used for training and the rest for testing. Extra 10,000 photos are applied to enlarge the gallery.

The NJU-ID database includes 256 high-resolution live photos with corresponding low-resolution ID photos. We follow the standard protocol in [63] to evaluate our proposed soft semantic representation on matching faces between live photo domain and ID photo domain.

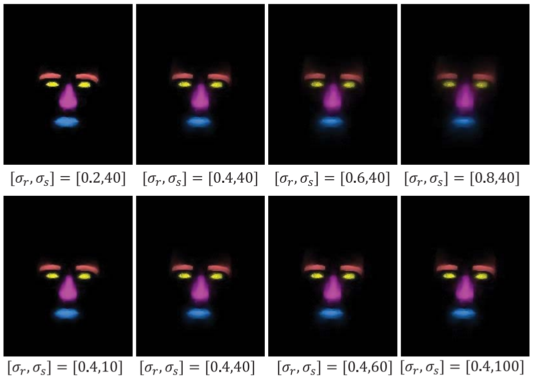
The CASIA NIR-VIS 2.0 database contains 725 subjects that are organized into two views. View 1 is used for parameter tuning and view 2 is used for performance evaluation. For each subject, there are multiple face images from the VIS domain as well as the NIR domain. We report the accuracy on the view 2 under the official protocol defined in [64].

In this article, we follow the same experimental protocols provided by the databases if they are available, such as the UoM-SGFS database [1] and the CASIA NIR-VIS 2.0 database [64]. Otherwise, we follow the same partition protocols given by corresponding papers. For e-PRIP database [41] and PRIP-VSGC database [62], we follow the partition protocol given by [41] to use 48 subjects for training and 75 subjects for testing. For the forensic sketch database [7], we follow the experimental protocol as [7] where 106 subjects are randomly selected for training and the rest 53 subjects for testing. For the NJU-ID database [63], we follow the exact 10-fold cross-validation protocol in [63] by randomly divide the database into 10 folds. It should be noticed that we usually conduct 10 random splits and average accuracies are reported on these databases without specific notification.

A. Implementation Details

Here we discuss the implementation details and parameter settings of our proposed algorithm on the e-PRIP database. All face images used in this article are firstly aligned and cropped into the size of 200 × 250. In the soft semantic face parsing phase, the parameters σ_r and σ_s are set to 0.4 and 40 respectively. We will analyze the influence of these two parameters in the following section. In soft semantic representation extraction, the size of image patches used for extracting CNN features is set to 32 × 32. In the soft content feature extraction stage, the keypoints are sampled with density of 10 × 10 grid. The influence of these two parameters in soft semantic feature extraction phase will also be discussed below. We simply follow [59] to take 4 × 4 image patches around each keypoint to explore facial character around the keypoint. The image patch level CNN is firstly trained with the Brown dataset [60]. We then perform fine-tuning on the cross-domain databases respectively. The hard samples chosen strategy [67] and data augmentation with random rotation are applied. Since the main contribution of this article is to apply the proposed soft semantic representation strategy for cross-domain face recognition, we choose to use an ordinary CNN instead of a well-designed complex module to better demonstrate the contributions of our method. In our experiments, the features extracted from the image patch level CNN can achieve satisfying performance on cross-domain face recognition task. In fact, the image patch level CNN has been presented in the task of image matching [68]–[71] to show great potentiality and effectiveness. The domain gap in our task can also be eliminated through the image patch level CNN under our network training strategy, which conducts fine-tuning on the cross-domain databases with hard samples chosen strategy [67] and data augmentation with random rotation. In the future, we will investigate the cooperation of our soft semantic parsing with more advanced deep networks. We train the network using a triplet loss and stochastic gradient descent optimization algorithm, with a starting learning rate of 0.1, momentum of 0.9 and weight decay of 0.0001. Except for following standard protocols, we will provide the performance under tenfold cross-validation in the remaining part of our article.

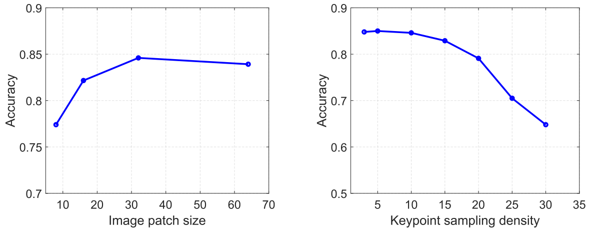
1) Influence of the Soft Semantic Parsing Parameters: In our proposed soft semantic face parsing algorithm, there are two parameters σ_r and σ_s which are needed to be determined. We provide a visual comparison about the influence of these two parameters in Fig. 4. From the figure we can see that our proposed soft semantic parsing algorithm can effectively smooth the unclear facial component boundaries while preserve the strong edge structure if the boundaries in the input face image is clear. Furthermore, we first fix σ_s at 40 as shown in the first row of Fig. 4. When σ_r is small, no smoothing is performed around the facial component boundaries. However, when σ_r is large, the soft semantic parsing map becomes extremely smooth. We then fix σ_r at 0.4 as shown in the second



$[\sigma_r, \sigma_s] = [0.2{,}40]$ $[\sigma_r, \sigma_s] = [0.4{,}40]$ $[\sigma_r, \sigma_s] = [0.6{,}40]$ $[\sigma_r, \sigma_s] = [0.8{,}40]$

$[\sigma_r, \sigma_s] = [0.4{,}10]$ $[\sigma_r, \sigma_s] = [0.4{,}40]$ $[\sigma_r, \sigma_s] = [0.4{,}60]$ $[\sigma_r, \sigma_s] = [0.4{,}100]$

Fig. 4. Influence of the soft semantic parsing parameters σ_r and σ_s. In the first row, σ_S.



(a) (b)

Fig. 5. Left subfigure shows the cross-domain face recognition accuracies with different image patch size; right subfigure shows the performance with different keypoint sampling densities in the soft content feature F_SC extraction.

row of the figure. When σ_s is small, no smoothing operation is performed. When σ_s becomes very large, the result is still not over smoothed. Therefore, we set σ_r and σ_s to be 0.4 and 40 respectively in this article, which could provide the best performance.

2) Influence of the Soft Semantic Representation Extraction Parameters: When applying the proposed soft semantic face parsing algorithm for cross-domain face recognition, we need to extract CNN features of image patches sampled from the cross-domain face images. Because we utilize the image patch level CNN for feature extraction, it is reasonable that the performance is related to the setting of image patch size. We evaluate our proposed method when utilizing different image patch sizes for feature extraction, as shown in the left subfigure of Fig. 5. When the image patch size is small, it is difficult for CNN to extract discriminative information from small image patches. Thus the performance is poor if we choose image patch size as 8×8 or 16×16 as shown in Fig. 5. With the increase of the image patch size, more meaningful and discriminative information together with semantic clue are taken into consideration, which boosts the cross-domain face recognition performance. However, the computation cost for image patch level CNN will increase with the larger image patch size. Therefore, we choose the image patch size to 32×32 as the default setting in this article.

Another parameter here is the keypoint sampling density, which controls the density of sampling keypoints within facial components to extract image patches for recognition. In the soft content feature extraction phase, we densely sample keypoints within the region of facial components and extract image patches around these keypoints for deep feature extraction. Therefore we evaluate the keypoint sampling strategy with different densities, as shown in the right subfigure of Fig. 5. When the density is small, the image patches will be sampled through a dense grid, and the computation cost will increase as well. Therefore, the keypoints are sampled with a reasonable density of 10×10, which yields satisfactory performance. For different cross-domain face recognition tasks, the influence of these parameters are similar. In our experiments, we find that applying the default settings to multiple datasets can achieve promising performance, thus proves the generalization ability of our method.

B. Experimental Results

1) Photo Domain vs. Sketch Domain: We first evaluate our proposed algorithm on matching face images between photo domain and sketch domain. Since the e-PRIP database and PRIP-VSGC database share the same 123 photos, we first provide the experimental comparison on these two databases in Table I. The sketches in e-PRIP and PRIP-VSGC are created with two different composite sketch generation software, in which the PRIP-VSGC is more difficult. Rank-10 accuracies after ten-fold cross-validation are reported. As shown in the table, the hand-crafted approaches, such as Fisherface, MCWLD and SSD, perform poor on this task. A composite components-based face sketch recognition was carefully designed by extracting hand-crafted features from components, which achieved rank-10 accuracies of 70.10% and 61.60% on e-PRIP and PRIP-VSGC respectively. The DEEPS method which combined a very deep CNN with morphed faces and transfer learning achieved a rank-10 accuracies of 80.80% and 54.90% here. Recently, the attribute information was introduced to sketch-photo face recognition in [72], which aimed to exploit the facial attributes to increase inter-personal variations. The attribute-assisted deep CNN (AADCNN) achieved rank-10 accuracies of 76.40% on e-PRIP and 72.30% on PRIP-VSGC. Our proposed soft semantic representation based method performs better than existing methods with the highest rank-10 accuracies of 84.60% and 77.12% on these two composite sketch databases respectively. The rationale behind our high performance is that the composite sketches are created by combining facial components in the software, which can be regarded as utilizing semantic clue during the composite sketch generation. Therefore, our method could fully exploit the semantic information among cross-domain faces, thus outperforms the state-of-the-art methods here.

The second experiment is conducted on another photo-sketch database, *i.e.* UoM-SGFS database. This is a large scale composite sketch database compared to e-PRIP and PRIP-VSGC. The sketches in this database is also created by combing semantic component parts together in the software. As shown in Table II, existing methods performed poor on this challenging large scale composite sketch database. The DLFace method extracted deep local descriptors around the facial landmarks, which is similar to our contour level

TABLE I
RECOGNITION PERFORMANCE (RANK-10, %) BETWEEN PHOTO DOMAIN AND COMPOSITE SKETCH DOMAIN ON THE E-PRIP DATABASE AND PRIP-VSGC DATABASE

Method	Accuracy on e-PRIP	Accuracy on PRIP-VSGC
Fisherface [61]	35.30	21.90
MCWLD [35]	24.00	15.40
SSD [73]	53.30	45.30
Composite-Components [74]	70.10	61.60
TLDNN [41]	60.20	52.00
CNN [75]	65.60	51.50
DEEPS [1]	80.80	54.90
DLFace [59]	82.80	74.60
AADCNN [72]	76.40	72.30
Proposed	84.60	77.12

TABLE II
RECOGNITION PERFORMANCE (%) BETWEEN PHOTO DOMAIN AND COMPOSITE SKETCH DOMAIN ON THE UOM-SGFS DATABASE

Method	Accuracy on subset A		Accuracy on subset B	
	Rank-1	Rank-10	Rank-1	Rank-10
PCA [61]	2.80	8.40	5.33	9.87
HAOG [76]	13.60	37.33	21.60	42.27
CBR [22]	5.73	18.80	7.60	25.47
LGMS [77]	21.87	51.20	43.47	73.60
VGG-Face [78]	9.33	31.07	16.13	48.00
D-RS [37]	22.13	49.33	40.80	70.80
D-RS+CBR [62]	25.87	56.00	42.93	75.87
DEEPS [1]	31.60	66.13	52.17	82.67
DLFace [59]	64.80	92.13	72.53	94.80
Proposed	70.16	94.60	73.83	95.10

feature extraction on initial hard face parsing map. Therefore, the DLFace method achieved a relatively high performance, with rank-1 accuracies of 64.80% and 72.53% on the subset A and subset B respectively. However, our proposed method adopts the soft face parsing map, where the unclear boundaries in input face image are smoothed while the clear boundaries in input face are kept. With the further help of the hierarchical soft semantic representation framework, our method achieved higher performance of 70.16% and 73.83% on this challenging database.

The third experiment here is conducted on a collected forensic sketch database, where the mug shot photos and sketches are collected from real law enforcement cases. In order to mimic real-world law enforcement scenario, the gallery is enlarged with 10,000 photos here. We report rank-50 accuracy after ten-fold cross-validation in Table III. Traditional face recognition algorithms including PCA and Fisherface performed extremely poor on this real-world cross-domain database. Instead of using single CNN architecture, a two-branch network architecture named residual compensation network (RCN) [51] was designed to learn separate deep features for cross-domain face images. A residual compensation part was integrated into CNN together with a modality discrepancy loss. This competitive RCN-10 achieves a rank-50 accuracy of 62.26% on the challenging forensic sketch dataset. DLFace [59] is also a competitive method that developed a deep local descriptor for cross-domain face recognition. It achieved a rank-50 accuracy of 57.64% here. Our method considers the soft semantic clue in the face parsing map and extracts representations from a hierarchical framework, thus achieves a rank-50 accuracy of 65.12%, which further demonstrate the advantage of the proposed method.

2) Live Photo Domain vs. ID Photo Domain: Experiment in this subsection shows the effectiveness of the proposed method on matching live photos with ID photos. The live photos are usually captured with a high resolution, varying illumination and complex background. On the contrary, the ID photos are often obtained under controlled environment with

TABLE III

RECOGNITION PERFORMANCE (RANK-50, %) BETWEEN MUGSHOT PHOTO AND FORENSIC SKETCH DOMAIN ON FORENSIC SKETCH DATABASE

Method	Accuracy	Method	Accuracy
PCA [61]	12.68	Fisherface [61]	17.14
D-RS [37]	20.80	VGG-Face [78]	24.46
DLFace [59]	57.64	G-HFR [7]	31.96
RCN-10 [51]	62.26	Proposed	65.12

TABLE IV

RECOGNITION PERFORMANCE (RANK-1, %) BETWEEN LIVE PHOTO DOMAIN AND ID PHOTO DOMAIN ON NJU-ID DATABASE

Method	Accuracy	Method	Accuracy
PCA [61]	23.00	CSR [79]	29.30
CCA [80]	20.30	CDFE [32]	24.60
MvDA [33]	16.50	CMML [81]	20.30
HMLCR [82]	27.70	CML [6]	30.90
DLFace [59]	43.46	Proposed	45.33

a low resolution. Therefore, face recognition across the live photo domain and the ID photo domain is quite challenging. We demonstrate rank-1 accuracies of existing method and the proposed method in Table IV. The hand-crafted methods achieve low accuracies on this task. The metric learning based method such as CML achieved rank-1 accuracy of 30.90%. The DLFace [59] is also competitive against the proposed method on other databases. It achieved rank-1 accuracy of 43.46% on the NJU-ID database. Our proposed method fully exploits the semantic clue in this scenario, which can be robust to varying resolution and illumination. Therefore, a highest accuracy of 45.33% is achieved on this challenging experiment, which illustrates the superiority of our method.

3) VIS Domain vs. NIR Domain: At last, we present the experiment on matching NIR images with VIS images in Table V. The NIR and VIS images are obtained by different sensors. Experiments in this subsection follow the protocols provided in CASIA NIR-VIS 2.0, which is well-define. The averaged accuracies in View 2 of the protocol are reported here. This is the largest publicly available NIR-VIS face database, which contains 725 subjects. The hand-crafted methods performed poor on this database. For example, the coupled spectral regression method [79] and the kernel prototype similarity based method [37] only achieved rank-1 accuracies of 33.8% and 28.2% respectively. Benefiting from the large scale dataset for network training, deep learning based methods can achieve high accuracies. Reference [44] utilized the high level features from deep CNN as domain specific units for heterogeneous face recognition, and achieved a rank-1 accuracy of 96.3%. The two-branch residual compensation network based on Resnet-10 backbone (RCN-10) [51] achieved a high accuracy of 99.32% on this large database. Deng *et al.* [47] published mutual component convolutional neural networks (MC-CNN) for cross-domain face recognition and also obtained promising performance of 99.39%, which incorporated their mutual component analysis into deep convolutional networks via regarding it as a fully-connected layer. Recently, He *et al.* [55] proposed to firstly synthesize VIS images from NIR faces through an adversarial cross-spectral face completion (CFC) framework, and then applied a light CNN as the baseline recognizer for cross-domain face recognition. In order to demonstrate the effectiveness of our method and to compare with state-of-the-art methods, we also evaluate our soft semantic representation on this database, and achieve rank-1 accuracy of 99.30% with the help of soft face parsing and hierarchical soft semantic representation extraction. The experimental result on CASIA NIR-VIS 2.0 database demonstrates the effectiveness and robustness of the proposed method on dealing with multiple cross-domain face recognition scenarios, thus presents the superiority of this article.

TABLE V

RECOGNITION PERFORMANCE (RANK-1, %) BETWEEN NIR DOMAIN AND VIS DOMAIN ON CASIA NIR-VIS 2.0 DATABASE

Method	Accuracy	Method	Accuracy
CSR [79]	33.80	D-RS [37]	28.20
CEFD [83]	85.60	TRIVET [84]	95.70
VGG-Face [78]	62.09	CNN [75]	85.90
IDNet [85]	87.10	SeetaFace [86]	68.03
CenterLoss [87]	87.69	LightCNN [88]	91.88
DSU [44]	96.30	IDR-128 [89]	97.33
WCNN [10]	98.70	DLFace [59]	98.68
RCN-10 [51]	99.32	MC-CNN [47]	99.39
CFC [55]	99.50	Proposed	99.30

C. Discussion

At last, we conduct an ablation study in this subsection to demonstrate the effectiveness of each part in the hierarchical soft semantic representation extraction phase. The motivation of our method is that the boundaries of facial components should be clear if they are visually clear in the face image, while the boundaries need to be smoothed if they are visually uncertain. We find that it is the best way to extract features directly from the soft semantic parsing maps for face representation in our experiments. The rationale is that the network could pay more attention on the facial component boundaries, while the texture information can be extracted from another contour level feature extraction branch. Han *et al.* [22] cropped facial components into rectangular patches and extract features from them, where the semantic shape characteristic and soft facial component boundaries were ignored for cross-domain face recognition. Table VI shows the performance of each part in the proposed hierarchical feature extraction framework, where F_{SS} represents directly matching the semantic shape characteristic between cross-domain faces. Since the shape

TABLE VI

ABLATION STUDY (RANK-10, %) FOR OUR SOFT SEMANTIC REPRESENTATION ON THE E-PRIP COMPOSITE SKETCH DATABASE

Method			Accuracy
F_{SS}	F_{SC}	F_{HC}	
√	-	-	26.44
-	√	-	78.24
√	√	-	81.60
-	-	√	82.80
-	√	√	84.12
√	√	√	84.60

features are compared directly without knowing any texture information, the performance of only using F_{SS} is poor. However, we find that fusing F_{SS} with the soft content feature F_{SC} and the contour level texture feature F_{HS} could achieve a promising result. It is reasonable because these features can capture the shared clues of cross-domain face images from different aspects. F_{SS} and F_{SC} focus on the semantic level information while F_{HC} is utilized to extract contour level texture information. Actually, [22] has found that fusion of several components which performed poor could achieve a high recognition rate. As shown in Table VI, the soft shape feature F_{SS} contributes slightly to the recognition accuracy when integrated with the soft content feature F_{SC}. The contour level feature F_{HC} extracted based on initial hard semantic parsing map achieved a similar performance compared to the soft semantic level features $F_{SS} + F_{SC}$. It can be noticed that either F_{SC} or F_{HC} could achieve a much better recognition performance than using the F_{SS} feature. The main reason is that both F_{SC} and F_{HC} are using patch level CNN for feature extraction, where the semantic appearance information and the facial contour texture information are fully exploited for cross-domain face recognition. We further conduct an experiment to fuse F_{SC} with F_{HC}, which could achieve better performance owing to the rationale that they are complementary. Considering that the contour level feature from the holistic face can be complementary to the soft semantic level features, we combine them together to form the set of our soft semantic representation, which achieves the highest performance.

In order to better illustrate the motivation of using semantic information, we will discuss our contribution from the following aspects: theoretical support from the creation process of cross-domain facial images, experimental support of an additional comparison experiment, and visual demonstration of the semantic parsing results on sketches and NIR images. First, we mainly focus on cross-domain face recognition task that includes matching visible photos with composite or forensic sketches. The composite sketches are usually created by software such as FACES [90] and Identi-Kit [91]. The software asks the user to select proper facial components from its template database and combines the selected components together to form the composite sketches. Therefore, semantic information plays an important role during the creation of the composite sketches. The forensic sketches are usually drawn by the forensic artists based on the description from the eyewitness. As indicated by the forensic sketch artist Gibson [92] and Taylor [93], it is easier for the eyewitness to describe the semantic facial component features of the suspect. This is because of the imperfect communication between the eyewitness and forensic artist, which makes the semantic information being a vital clue for creating the forensic sketches. The NIR face images are captured through near-infrared sensors under poor lighting environments [64]. Because the texture information in NIR faces are missing, the remaining semantic shape information is an important clue for NIR face recognition, which is never exploited before. Because there is no exaggeration in terms of shape during the creation process of sketches or NIR face images, the semantic information is an important bridge between face images in these different domains. Therefore, we develop a hierarchical soft semantic representation framework that combines the features from densely sampled contour level and the soft semantic component level to improve the performance of cross-domain face recognition. Next, we conduct an additional experiment to show what the performance of the proposed method would be if the facial components are replaced by regular rectangular facial blocks. We perform this comparison experiment on the e-PRIP database. By dividing the face images into regular 8×8 subregions and 16×16 subregions, the rank-10 accuracies of cross-domain face recognition on e-PRIP are 81.20% and 82.40% respectively. Our method achieves a rank-10 accuracy of 84.60% on this database, which supports the contribution of our method. Finally, we present several visual demonstration of the soft semantic parsing results on composite sketches and NIR face images, as shown in Fig. 6. As described above, the composite sketches are created by combining semantic facial components from the template database in the software, there is little distortion in these sketches. As for the NIR images, parts of the facial contours are not clear due to the poor quality of NIR sensors. Therefore, the boundaries of these unclear facial contours will be smoothed in the parsing map guided by the NIR images, which fits the motivation of our soft parsing in comparison to hard parsing. Therefore, the semantic information can be well exploited by our soft parsing algorithm to contribute to face recognition.

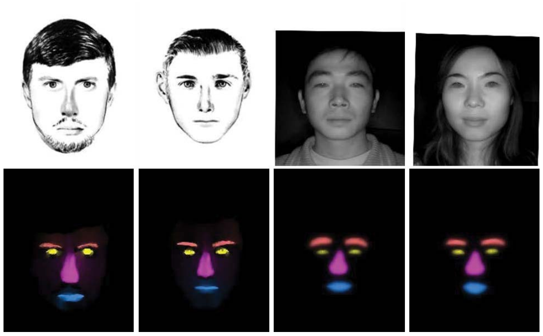


Fig. 6. Visual demonstration of the soft semantic parsing results on composite sketches and NIR face images.

As introduced in section III-A, our proposed soft semantic face parsing is motivated by the domain transform method proposed in [57]. However, we have made a number of improvements in order to firstly apply edge-preserving filtering for cross-domain face representation. *Firstly*, we improve the edge-preserving filtering in [57] into a discrete version, since we need to process the face image on pixel level. *Secondly*, in [57] the edge-preserving filter is mainly applied to process the signal itself. We find that simply smoothing the face image may lose discriminative information for recognition. Therefore, we take the face image as the guidance of the filtering process for the semantic parsing map image. Specifically, we detect the spatial edge information from the face image, and then smooth the facial components boundaries in the parsing map image. Our assumption is simple that the initial hard parsing map is inaccurate since there are visually unclear boundaries in the face image, such as the bridge of the nose. The major contribution of our method is the proposed soft semantic face parsing method to help improve the performance of cross-domain face recognition. As noted in section III, the guided edge-preserving filter used in our soft semantic face parsing is motivated by [57]. However, we have made substantial modifications based on [57]so that the soft semantic information in face images can be exploited for recognition. Actually, the edge-preserving filter proposed in [57] was designed for natural image-to-image domain transform task regardless of detailed semantic information for recognition purpose. In our method, we focus on facial images with regular edges and facial structures. Our motivation is to fully exploit and character the facial contours in semantic parsing map based on a soft parsing manner. By modifying the method in [57] to a discrete version, we propose to consider the original face image as the guidance of the edge-preserving filter process. In our method, the visually clear boundaries will be kept while the visually uncertain boundaries will be smoothed. Our contribution lies in that we firstly apply this original face image guided edge-preserving filter into cross-domain face recognition scenario, and we observe that the recognition performance is benefited from the spatial soft semantic information, which is never explored before. Therefore, we firstly apply the guided edge-preserving filter for cross-domain face recognition task, where the spatial semantic information is taken into consideration in face representation extraction process. We further develop a hierarchical soft semantic representation framework to fully exploit the semantic clue for cross-domain face recognition. Experimental results demonstrate the outperforming performance of the proposed soft semantic parsing method, which can prove our assumption as well as the novelty and superiority of this article.

Furthermore, we also compare and visualize the performance of our proposed soft semantic parsing with two other strategies (using holistic face without semantic parsing and initial hard semantic parsing) in Fig. 7. The w/o semantic parsing (holistic face) refers to extract image patches around the keypoints densely sampled from the whole face image. The w/ semantic parsing (hard) denotes to a similar framework introduced in this article where the soft semantic parsing map is replaced with the initial hard semantic map. The last w/ semantic parsing (soft) represents the proposed method. From the experimental comparison, it can be seen that the semantic information can be helpful to the cross-domain face recognition task. Furthermore, our proposed soft face parsing can better exploit the semantic clue to further improve the performance.

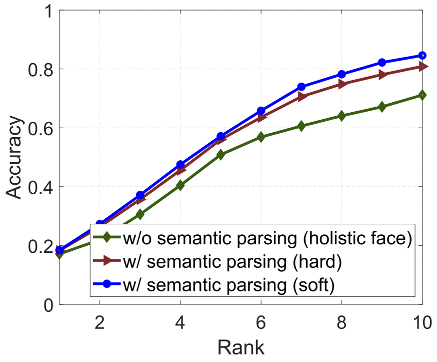


Fig. 7. Illustration about the effectiveness of the proposed soft semantic parsing.

V. Conclusion

Because of the huge domain discrepancy, cross-domain face recognition is a challenging problem. In this article, we present a novel soft semantic representation for cross-domain face recognition. A soft face parsing algorithm is firstly introduced, which can generate a soft semantic parsing map under the guidance of the input face image to avoid inaccurate segmentation around component boundaries in initial hard parsing map. A hierarchical soft semantic representation extraction framework is then presented, where soft shape feature, soft content feature, and hard semantic level contour feature form the set of soft semantic representation. By fusing the set of soft semantic representation, the proposed method achieves the best performance on multiple cross-domain face databases.

Future work includes applying the proposed soft semantic parsing algorithm to other fields. We emphasize the proposed soft semantic face parsing in a separate subsection in this article since it can also be regarded as a pre-processing phase for face recognition. The inaccurate segmentation of unclear visual boundaries around facial components can be smoothed, which is helpful for better cross-domain facial representation. We believe that the characteristics of smoothing the inaccurate segmentation around unclear boundaries while preserving the edges at clear boundaries of face image, it can be also integrated into other related tasks such as face deblurring, face denoising and face systhesis.

References

[1] C. Galea and R. A. Farrugia, “Matching software-generated sketches to face photographs with a very deep CNN, morphed faces, and transfer learning,” *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 6, pp. 1421–1431, Jun. 2018.

[2] Y. Shi and A. K. Jain, “DocFace: Matching ID document photos to selfies,” in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst.*, Oct. 2018, pp. 56–67.

[3] H. Zhang, H. Han, J. Cui, S. Shan, and X. Chen, “RGB-D face recognition via deep complementary and common feature learning,” in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2018, pp. 1–8.

[4] S. Ouyang, T. Hospedales, Y.-Z. Song, X. Li, C. C. Loy, and X. Wang, "A survey on heterogeneous face recognition: Sketch, infra-red, 3D and low-resolution," *Image Vis. Comput.*, vol. 56, pp. 28–48, Dec. 2016.

[5] 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.

[6] J. Huo, Y. Gao, Y. Shi, and H. Yin, "Cross-modal metric learning for AUC optimization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4844–4856, Oct. 2018.

[7] C. Peng, X. Gao, N. Wang, and J. Li, "Graphical representation for heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 301–312, Feb. 2017.

[8] L. Zhang, L. Lin, X. Wu, S. Ding, and L. Zhang, "End-to-end photo-sketch generation via fully convolutional representation learning," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2015, pp. 627–634.

[9] D. Yi, Z. Lei, and S. Z. Li, "Shared representation learning for heterogenous face recognition," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, May 2015, pp. 1–7.

[10] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein CNN: Learning invariant features for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1761–1773, Jul. 2019.

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

[12] S. Liu, J. Yang, C. Huang, and M.-H. Yang, "Multi-objective convolutional learning for face labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3451–3459.

[13] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, and M.-H. Yang, "Deep semantic face deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8260–8269.

[14] Y. Song *et al.*, "Joint face hallucination and deblurring via structure generation and detail enhancement," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 785–800, Jun. 2019.

[15] R. Wan, B. Shi, H. Li, L.-Y. Duan, and A. C. Kot, "Face image reflection removal," 2019, *arXiv:1903.00865*. [Online]. Available: http://arxiv.org/abs/1903.00865

[16] D. Zhang, L. Lin, T. Chen, X. Wu, W. Tan, and E. Izquierdo, "Content-adaptive sketch portrait generation by decompositional representation learning," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 328–339, Jan. 2017.

[17] S. Zhang, R. Ji, J. Hu, Y. Gao, and C.-W. Lin, "Robust face sketch synthesis via generative adversarial fusion of priors and parametric sigmoid," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1163–1169.

[18] J. Yu *et al.*, "Towards realistic face photo-sketch synthesis via composition-aided GANs," 2017, *arXiv:1712.00899*. [Online]. Available: http://arxiv.org/abs/1712.00899

[19] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1062–1071.

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[21] A. Zadeh, T. Baltrusaitis, and L.-P. Morency, "Convolutional experts constrained local model for facial landmark detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jul. 2017, pp. 2519–2528.

[22] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain, "Matching composite sketches to face photos: A component-based approach," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 1, pp. 191–204, Jan. 2013.

[23] C. Chen and A. Ross, "Matching thermal to visible face images using a semantic-guided generative adversarial network," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2019, pp. 1–8.

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

[25] N. Wang, J. Li, D. Tao, X. Li, and X. Gao, "Heterogeneous image transformation," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 77–84, Jan. 2013.

[26] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, and J. Li, "Multiple representations-based face sketch–photo synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2201–2215, Nov. 2016.

[27] 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.

[28] 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.

[29] N. Wang, M. Zhu, J. Li, B. Song, and Z. Li, "Data-driven vs. model-driven: Fast face sketch synthesis," *Neurocomputing*, vol. 257, pp. 214–221, Sep. 2017.

[30] N. Wang, X. Gao, L. Sun, and J. Li, "Anchored neighborhood index for face sketch synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2154–2163, Sep. 2018.

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

[32] D. Lin and X. Tang, "Inter-modality face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 13–26.

[33] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 808–821.

[34] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Li, "Heterogeneous face recognition from local structures of normalized appearance," in *Proc. IAPR Int. Conf. Biometrics*, 2009, pp. 209–218.

[35] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetically optimized MCWLD for matching sketches with digital face images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1522–1535, Oct. 2012.

[36] H. Kiani Galoogahi and T. Sim, "Face sketch recognition by local radon binary pattern: LRBP," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1837–1840.

[37] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013.

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

[39] C. Peng, X. Gao, N. Wang, and J. Li, "Face recognition from multiple stylistic sketches: Scenarios, datasets, and evaluation," *Pattern Recognit.*, vol. 84, pp. 262–272, Dec. 2018.

[40] Y. Jin, J. Lu, and Q. Ruan, "Coupled discriminative feature learning for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 3, pp. 640–652, Mar. 2015.

[41] M. Paritosh, M. Vasta, and R. Singh, "Composite sketch recognition via deep network-a transfer learning approach," in *Proc. IAPR Int. Conf. Biometrics*, 2015, pp. 251–256.

[42] C. Galea and R. A. Farrugia, "Forensic face photo-sketch recognition using a deep learning-based architecture," *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1586–1590, Nov. 2017.

[43] C. Galea and R. A. Farrugia, "Face photo-sketch recognition with log-Gabor filters and statistical correlation coefficients," in *Proc. Eur. Signal Process. Conf.*, 2016, pp. 2240–2244.

[44] T. Pereira, A. Anjos, and S. Marcel, "Heterogeneous face recognition using domain specific units," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 7, pp. 1803–1816, Jul. 2018.

[45] H. Li, H. Hu, and C. Yip, "Age-related factor guided joint task modeling convolutional neural network for cross-age face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2383–2392, Sep. 2018.

[46] S. P. Mudunuri, S. Venkataramanan, and S. Biswas, "Dictionary alignment with re-ranking for low-resolution NIR-VIS face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 4, pp. 886–896, Apr. 2019.

[47] Z. Deng, X. Peng, Z. Li, and Y. Qiao, "Mutual component convolutional neural networks for heterogeneous face recognition," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3102–3114, Jun. 2019.

[48] S. Mehdi Iranmanesh and N. M. Nasrabadi, "Attribute-guided deep polarimetric thermal-to-visible face recognition," 2019, *arXiv:1907.11980*. [Online]. Available: http://arxiv.org/abs/1907.11980

[49] V. Talreja, F. Taherkhani, M C Valenti, and N. M. Nasrabadi, "Attribute-guided coupled GAN for cross-resolution face recognition," 2019, *arXiv:1908.01790*. [Online]. Available: http://arxiv.org/abs/1908.01790

[50] X. Zhu *et al.*, "Large-scale bisample learning on ID versus spot face recognition," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 684–700, Jun. 2019.

[51] Z. Deng, X. Peng, and Y. Qiao, "Residual compensation networks for heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8239–8246.

[52] S. Yu, H. Han, S. Shan, A. Dantcheva, and X. Chen, "Improving face sketch recognition via adversarial sketch-photo transformation," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2019, pp. 1–8.

[53] X. Wu, L. Song, R. He, and T. Tan, "Coupled deep learning for heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1679–1686.

[54] J. Yu, J. Cao, Y. Li, X. Jia, and R. He, "Pose-preserving cross-spectral face hallucination," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 1018–1024.

[55] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, “Adversarial cross-spectral face completion for NIR-VIS face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1025–1037, May 2020.

[56] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, “Dual variational generation for low-shot heterogeneous face recognition,” in *Proc. 33rd Conf. Neural Inf. Process. Syst.*, 2019, pp. 2674–2683.

[57] E. Gastal and M. M. Oliveira, “Domain transform for edge-aware image and video processing,” *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–12, 2011.

[58] W. Brendel and M. Bethge, “Approximating CNNs with bag-of-local features models works surprisingly well on ImageNet,” in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–15.

[59] C. Peng, N. Wang, J. Li, and X. Gao, “DLFace: Deep local descriptor for cross-modality face recognition,” *Pattern Recognit.*, vol. 90, pp. 161–171, Jun. 2019.

[60] M. Brown, G. Hua, and S. Winder, “Discriminative learning of local image descriptors,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 43–57, Jan. 2011.

[61] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[62] S. J. Klum, H. Han, B. F. Klare, and A. K. Jain, “The FaceSketchID system: Matching facial composites to mugshots,” *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2248–2263, Dec. 2014.

[63] J. Huo, Y. Gao, Y. Shi, W. Yang, and H. Yin, “Ensemble of sparse cross-modal metrics for heterogeneous face recognition,” in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 1405–1414.

[64] S. Z. Li, D. Yi, Z. Lei, and S. Liao, “The CASIA NIR-VIS 2.0 face database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 348–353.

[65] A. Martinez and R. Benavente, “The AR face database,” CVC, Barcelona, Spain, Tech. Rep. #24, 1998.

[66] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, “The FERET evaluation methodology for face-recognition algorithms,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[67] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[68] S. Zagoruyko and N. Komodakis, “Learning to compare image patches via convolutional neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4353–4361.

[69] B. G. Vijay Kumar, G. Carneiro, and I. Reid, “Learning local image descriptors with deep Siamese and triplet convolutional networks by minimizing global loss functions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5385–5394.

[70] T.-Y. Yang, J.-H. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, “DeepCD: Learning deep complementary descriptors for patch representations,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3314–3322.

[71] X. Wei, Y. Zhang, Y. Gong, and N. Zheng, “Kernelized subspace pooling for deep local descriptors,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1867–1875.

[72] S. M. Iranmanesh, H. Kazemi, S. Soleymani, A. Dabouei, and N. M. Nasrabadi, “Deep sketch-photo face recognition assisted by facial attributes,” in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst.*, Oct. 2018, pp. 1–10.

[73] P. Mittal, A. Jain, G. Goswami, R. Singh, and M. Vatsa, “Recognizing composite sketches with digital face images via SSD dictionary,” in *Proc. IAPR Int. Conf. Biometrics*, 2014, pp. 1–6.

[74] D. Liu, J. Li, N. Wang, C. Peng, and X. Gao, “Composite components-based face sketch recognition,” *Neurocomputing*, vol. 302, pp. 46–54, Aug. 2018.

[75] S. Saxena and J. Verbeek, “Heterogeneous face recognition with CNNs,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–491.

[76] H. K. Galoogahi and T. Sim, “Inter-modality face sketch recognition,” in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 224–229.

[77] C. Galea and R. A. Farrugia, “Face photo-sketch recognition using local and global texture descriptors,” in *Proc. Eur. Signal Process. Conf.*, 2016, pp. 2240–2244.

[78] O. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *Proc. Brit. Mach. Vis. Conf.*, 2015, p. 6.

[79] Z. Lei and S. Z. Li, “Coupled spectral regression for matching heterogeneous faces,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1123–1128.

[80] H. Hotelling, “Relations between two sets of variate,” *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.

[81] A. Mignon and F. Jurie, “CMML: A new metric learning approach for cross modal matching,” in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 1–14.

[82] L. Wu *et al.*, “Heterogeneous metric learning with content-based regularization for software artifact retrieval,” in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 610–619.

[83] D. Gong, Z. Li, W. Huang, X. Li, and D. Tao, “Heterogeneous face recognition: A common encoding feature discriminant approach,” *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2079–2089, May 2017.

[84] X. Liu, L. Song, X. Wu, and T. Tan, “Transferring deep representation for NIR-VIS heterogeneous face recognition,” in *Proc. IAPR Int. Conf. Biometrics*, 2016, pp. 1–8.

[85] C. Reale, N. M. Nasrabadi, H. Kwon, and R. Chellappa, “Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 54–62.

[86] X. Liu, M. Kan, W. Wu, S. Shan, and X. Chen, “VIPLFaceNet: An open source deep face recognition SDK,” *Frontiers Comput. Sci.*, vol. 11, no. 2, pp. 208–218, Apr. 2017.

[87] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.

[88] X. Wu, R. He, Z. Sun, and T. Tan, “A light CNN for deep face representation with noisy labels,” 2015, *arXiv:1511.02683*. [Online]. Available: http://arxiv.org/abs/1511.02683

[89] R. He, X. Wu, and Z. Sun, “Learning invariant deep representation for NIR-VIS face recognition,” in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2000–2006.

[90] *FACES*. Accessed: Aug. 1, 2020. [Online]. Available: http://www.iqbiometrix.com

[91] *Identi-Kit*. Accessed: Aug. 1, 2020. [Online]. Available: http://www.identikit.net/

[92] L. Gibson, *Forensic Art Essentials: A Manual for Law Enforcement Artists*. Waltham, MA, USA: Academic, 2010.

[93] K. Taylor, *Forensic Art and Illustration*. Boca Raton, FL, USA: CRC Press, 2000.

Chunlei Peng (Member, IEEE) received the B.Sc. degree in electronic and information engineering from Xidian University, Xi’an, China, in 2012, and the Ph.D. degree in information and telecommunications engineering in 2017. From September 2016 to September 2017, he has been a visiting Ph.D. student with the Duke University, NC, USA. He is currently working with the School of Cyber Engineering, Xidian University. His current research interests include computer vision, pattern recognition, and machine learning.

Nannan Wang (Member, IEEE) received the B.Sc. degree in information and computation science from the Xi’an University of Posts and Telecommunications in 2009, and the Ph.D. degree in information and telecommunications engineering in 2015. From September 2011 to September 2013, he was a visiting Ph.D. student with The University of Technology, Sydney, NSW, Australia. He is currently working with the State Key Laboratory of Integrated Services Networks, Xidian University. His current research interests include computer vision, pattern recognition, and machine learning. He has published more than 50 articles in refereed journals and proceedings, including the *International Journal of Computer Vision*, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (T-NNLS), the IEEE TRANSACTIONS ON IMAGE PROCESSING (T-IP), and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (T-CSVT).

Jie Li received the B.Sc. degree in electronic engineering, the M.Sc. degree in signal and information processing, and the Ph.D. degree in circuit and systems from Xidian University, Xi'an, China, in 1995, 1998, and 2004, respectively. She is currently a Professor with the School of Electronic Engineering, Xidian University. Her research interests include image processing and machine learning. In these areas, she has published around 50 technical articles in refereed journals and proceedings, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (T-NNLS), the IEEE TRANSACTIONS ON IMAGE PROCESSING (T-IP), the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (T-CSVT), and *Information Sciences*.

Xinbo Gao (Senior Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Post-Doctoral Research Fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education, a Professor of *Pattern Recognition and Intelligent System*, and the Director of the State Key Laboratory of Integrated Services Networks, Xi'an. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications. He has published five books and around 200 technical articles in refereed journals and proceedings. He is on the Editorial Boards of several journals, including *Signal Processing* (Elsevier) and *Neurocomputing* (Elsevier). He has served as the General Chair/Co-Chair, the Program Committee Chair/Co-Chair, or a PC Member for around 30 major international conferences. He is also a fellow of the Institution of Engineering and Technology.