# Deep Attribute Guided Representation for Heterogeneous Face Recognition

**Decheng Liu[†], Nannan Wang[‡*], Chunlei Peng[§], Jie Li[†], Xinbo Gao[†]**
[†] State Key Laboratory of Integrated Services Networks,
School of Electronic Engineering, Xidian University, Xi'an 710071, China
[‡] State Key Laboratory of Integrated Services Networks,
School of Telecommunications Engineering, Xidian University, Xi'an 710071, China
[§] School of Cyber Engineering, Xidian University, Xi'an 710071, China

## Abstract

Heterogeneous face recognition (HFR) is a challenging problem in face recognition, subject to large texture and spatial structure differences of face images. Different from conventional face recognition in homogeneous environments, there exist many face images taken from different sources (including different sensors or different mechanisms) in reality. Motivated by human cognitive mechanism, we naturally utilize the explicit invariant semantic information (face attributes) to help address the gap of different modalities. Existing related face recognition methods mostly regard attributes as the high level feature integrated with other engineering features enhancing recognition performance, ignoring the inherent relationship between face attributes and identities. In this paper, we propose a novel deep attribute guided representation based heterogeneous face recognition method (DAG-HFR) without labeling attributes manually. Deep convolutional networks are employed to directly map face images in heterogeneous scenarios to a compact common space where distances mean similarities of pairs. An attribute guided triplet loss (AGTL) is designed to train an end-to-end HFR network which could effectively eliminate defects of incorrectly detected attributes. Extensive experiments on multiple heterogeneous scenarios (composite sketches, resident ID cards) demonstrate that the proposed method achieves superior performances compared with state-of-the-art methods.

## 1 Introduction

Face recognition is an important and challenge problem in computer vision. Recently great progress has been achieved, but there still exist many challenging face recognition scenarios. In the real world, face images are captured from different sources. The conventional homogeneous face recognition methods perform poorly applied in the different modalities of heterogeneous face images, such as visual images (VIS), face sketches (generated by hands or software), ID card photos (embedded in the card), near infrared images (captured through near infrared devices). For example, in the law enforcement agency, when no face photo image of the suspect is available or there are poor quality images in video surveillance, face sketches created by forensic artist or composite-generation software are utilized as the probe matching with gallery photos. With technological improvement, law enforcement agencies have started to utilize the generation software to produce composite sketches as a replacement of hand-drawn sketches. It should be noted that the face attributes information of the suspect could be acquired directly from the language description of eyewitness. In addition, when we utilize the resident ID card photos in the related identification or verification tasks, face attribute related information could also be acquired in the resident ID card directly. Naturally, we aim to integrate face attribute discriminative information to address the great gap between different modalities in heterogeneous face recognition.

Existing heterogeneous face recognition (HFR) methods could be classified into three categories: Feature descriptor based methods, Synthesis based methods and Common space projection based methods. Feature descriptor based methods aim to directly extract modality invariant features which measured for recognition. Synthesis based methods firstly transform images in one modality to another modality which would make these images in homogeneous scenarios, and then conventional homogeneous face recognition methods could be directly utilized. Common space methods attempt to project heterogeneous face images into a latent common space where the probe image and the gallery images could be matching directly. Here we give a brief review of representative HFR methods. **Synthesis based HFR methods:** [Tang and Wang, 2003] presented an Eigen-transform algorithm for sketch-photo synthesis. [Liu *et al.*, 2005] employed local linear embedding (LLE) to synthesize face sketches. Considering the relationship between face image patches and its neighboring patches, [Wang and Tang, 2009] exploited the Markov random field (MRF) for synthesis. Instead of directly matching synthesized sketches, [Peng

*Corresponding author: Nannan Wang (nnwang@xidian.edu.cn)

*et al.*, 2017] proposed a novel graphical representation (G-HFR) where Markov network are employed to represent image patches separately. Apparently the quality of synthesized sketches could influence the matching performance and the image synthesis process is a complex problem itself; **Feature descriptor based HFR methods:** [Klare *et al.*, Mar 2011] explored the multiple hand-crafted features and proposed a local feature based discriminant analysis. [Mittal *et al.*, 2015] presented a transfer learning based representation method. [Lu *et al.*, 2017] proposed an unsupervised feature learning method which learns features from raw pixels. However, this kind of methods would be utilized with high computational complexity; **Common space based HFR methods:** [Lin and Tang, 2006] firstly proposed a common discriminant feature extraction (CDFE) approach. A multi-view discriminant analysis (MvDA) method [Kan *et al.*, 2016] was proposed to exploit both inter-view and intra-view correlations of heterogeneous face images. [Huo *et al.*, 2017] proposed a margin based cross-modality metric learning to address the gap of different modalities. Yet the projection procedure may losses some discriminative information. To address the separation between heterogeneous face images, many methods proposed regard face attributes as supplementary discriminative information to enhance the matching performance. The proposed approach falls under the attribute related recognition approach with CNNs, but with several differences compared with existing methods [Kumar *et al.*, 2011] [Ouyang *et al.*, 2014] [Hu *et al.*, 2017]. Unlike most attribute related recognition methods [Ouyang *et al.*, 2014] [Hu *et al.*, 2017] [Li *et al.*, 2015], the proposed method could automatically evaluate attributes of face photos in training stage instead of manual labeling ; Unlike existing methods that focus on the feature level fusion method [Ouyang *et al.*, 2014] [Hu *et al.*, 2017] with multi-step frameworks, the proposed method focuses on directly integrating attribute discriminative information with an end-to-end structure framework to reduce the computational complexity; Unlike [Mittal *et al.*, 2017] attribute feedback only utilizes the most reliable attributes to filter ranked list in the testing stage, the proposed method doesn't need attributes labeled when matching and simultaneously eliminates the adverse effect of incorrect prediction.

This paper proposes a novel deep attribute guided representation for heterogeneous face recognition (DAG-HFR), which directly integrates the relationship between attributes and identities of different subjects. In real world scenario, some kinds of heterogeneous face images (e.g. composite sketches or resident ID card photos) are marked with attributes in the generation process. The attribute evaluation network is pretrained to detect attributes of face photos captured in the wild, and then the designed convolutional network could map heterogeneous face images into a common space to address the great gaps between two different modalities face images. Motivated by the empirical rules in human cognition mechanism, we propose the attribute guided triplet loss (AGTL) to eliminate the negative effect of wrong attribute prediction. Euclidean distances are utilized to measure similarities between the exploited representations of input images. Finally, similarity scores between the probe images and gallery images are calculated for matching. The main reason of superior performances is that we manually choose some discriminative facial attributes to enhance the recognition performance.

The main contributions of this paper are summarized as follows:

1. We employ a deep attribute guided representation for heterogeneous face recognition, which could effectively integrate face attributes discriminative information, and automatically detect face attributes of photo images without manual attributes labeling in the training stage.
2. The attribute guided triplet loss is designed by the inherent relationship of identities and face attributes, which can reduce the negative influence of incorrectly detected facial attributes, and make the representation more discriminative.
3. Experimental results illustrate the superior performance of proposed method compared with the state-of-the-art HFR methods. Meanwhile, the exploited representations for matching are only 128 dimensions which could be effectively used in large scale images matching tasks.

We organize the rest of this paper as follows. Section 1 gives a brief review of representative HFR methods. In Section 2, we present the deep attribute guided representation for heterogeneous face recognition. Section 3 shows the experimental results and related analysis, and the conclusion is drawn in Section 4.

## 2 Deep Attribute Guided Representation for Heterogeneous Face Recognition

In this section, we present a novel framework for HFR, which is called deep attribute guided representation for heterogeneous face recognition (DAG-HFR). It is noted that we take face sketch-photo recognition as an example to describe the proposed method for ease of representation, which could be easily generalized to other heterogeneous face recognition applications. In the subsections, we will introduce our motivation firstly, and then give a detailed explanation of our proposed DAG-HFR method.

### 2.1 Motivation

Motivated by the generation procedure of heterogeneous face images, we find that there indeed exists some important inherent and discriminative information in heterogeneous face recognition. Here we take face sketches as an example. In the real scene, the police could not capture direct photos from suspects. Instead, they could find eyewitness cooperating with a forensic artist or software operator to generate face sketches. It is the language description of eyewitness leads to the face sketches generation. And meanwhile these description sentences are composed of different kinds of facial attributes, such as gender, age, eyeglasses, etc. Naturally, these different kinds of facial attributes are marked without any extra labor in the generation procedure. Compared with generated face sketches, the facial attributes information derived from eyewitness' description should be more reliable, which could effectively address the communication gap [Ouyang *et al.*, 2016].

In daily life, when we try remember someone not seen for a long time, the first things come to our mind are not the concrete portraits but the attributes related information, like gender, ages, eyeglasses or not, beard or not, etc. Thus face attributes could be explicit and discriminative information naturally. Inspired from that, we naturally present an **empirical attribute guided rule** in identification task: For a pair of face images, the more similar between attributes of the pairs are, the more likely the same identity the pairs belong to. However, in most case the evaluation of facial attributes couldn't be completely correct. Therefore, directly utilizing the probe's attributes information to filter gallery photos may make more mistakes. Since we design the attribute guided triplet loss to make images of different identities more discriminative with different attributes. Experiments show the proposed method is robust to the facial attributes evaluation performance.

## 2.2 DAG-HFR

In this section, we present a novel attribute guided representation approach for heterogeneous face recognition. Without loss of generality and for ease of representation, we take face composite sketch-photo recognition as a representative example to describe the proposed method. Figure 1 shows the framework of the proposed method. Due to the limited heterogeneous face images sources, we make an attempt to construct more training data to improve the generalization of proposed model. Inspired by triplet loss in [Schroff *et al.*, 2015], we naturally combine lots of triplets with the probe sketches and the gallery photos heterogeneous face recognition task. Furthermore, we use the pre-trained Attribute Evaluation Network with multi-task architecture to detect face attributes for face photos in triplets. Attribute guided triplet loss (AGTL) are presented derived from the empirical attribute guided rule to eliminate the effect of incorrect attributes evaluations. The details are introduced as follows.

Motived by the successful application of triplet ranking loss [2015] on conventional face recognition related tasks, we design a novel attribute guided triplet loss here, which could effectively eliminate the negative influence of attribute prediction. Considering composite sketch-face pairs $\{(s_1, p_1), (s_2, p_2), ..., (s_N, p_N)\}$, $N$ is the number of subject. The proposed attribute guided triplet loss is defined as:

$$\begin{aligned} &L(s_1, ..., s_N, p_1^p, ..., p_N^p, p_1^n, ..., p_N^n, w) \\ &= \sum_i^N [\Phi(f(s_i), f(p_i^p)) - \Phi(f(s_i), f(p_i^n)) + \lambda \Psi(p_i^n)]_+, \end{aligned} \tag{1}$$

where

$$\Psi(p_i^n) = \left\| s_i^{attri} - g(p_i^n) \right\|_2 + m.$$

Here $(\cdot)_+$ is the same with $\max(0, \cdot)$.$\Phi(\cdot, \cdot)$ is the function to measure distances, we choose $\ell^2$ norm here.$f(\cdot)$ means the non-linear network mapping function, $s_i^{attri} \in \{(0, 1), (1, 0)\}$ means the probe sketch inherent binary attributes, $m$ is a margin. Here different from $f(\cdot)$ (mapping function) $g(\cdot)$ represents the similarity scores for binary attributes of input photo images. More details about $g(\cdot)$ are shown in the next paragraph. In the face sketch-photo recognition scenario, we aim to ensure that probe face sketch $s_i$ is closer to face photos $p_i^p$ (positive) of the same person than it is to any photos $p_i^n$ (negative) of any other subjects. For heterogeneous face recognition, given $s_i$ (anchor) of a subject, the gallery photo with the same identity is $p_i^p$ (positive), the rest gallery photos with different identities are $p_i^n$ (negative). Unlike the most conventional homogeneous face databases, most heterogeneous face databases contain pairs of face images with different modalities. The great gaps between different modalities may lead to the distance between the negative pairs (also called interpersonal distance) is larger than that between the positive pairs (also called intrapersonal distance) in original space. It is harmful for matching performance. When the proposed AGTL is applied, the intrapersonal distance is larger than the interpersonal distance for one subject. Besides, we follow the attribute guided rule illustrated before, and aim to push the negative photos $p_i^n$ detected different attributes with the sketch $s_i$ farther than the negative photos $p_i^n$ detected the same attribute with that. Thus the more discriminative attribute guided information could be exploited in HFR. The hyper-parameter $\lambda$ balances the importance of the attribute discriminability influence. Experiments results show the AGTL not only effective integrated attribute information, but also could eliminate the negative effect of incorrectly detected attributes.

Suppose we have $M$ facial attributes and the heterogeneous face training photos $p_i$ for the $m$-th attribute are separately denoted as $a_i^m$ . The proposed Attribute Evaluation Network can be formulated to minimize:

$$\arg\min_w \sum_{m=1}^{M} \sum_{n=1}^{N} L(a_i^m, g(p_i, w)), \tag{2}$$

where $g(p_i, w)$ means the prediction result with input image $p_i$ and the parameter of attribute evaluation network $w$ ; $a_i^m$ is the attribute labelled for $p_i$ ; $L(\cdot)$ is the loss function designed for prediction. Here we choose the cross entropy loss for two class classification. Finally, $g(\cdot)$ is used to predict face photos single binary attributes as shown in Figure 1. Here the evaluated similarity score of one face photo is a two-dimensions positive value vector with normalization. Euclidean distance measures the distances of face photos with different attributes predicted. It is noted that this pre-trained attribute evaluation network is only used to provide the attribute similarity scores for input face photos which doesn't need be updated in the training stage.

## 2.3 Implementation Details

We assume the training dataset with $N$ composite-sketch pairs. As illustrated before, $N \times (N-1)$triplets are naturally combined to pass through the network in training stage. Given $s_i$(anchor) is the sketch image here, we select the one photo image of the same subject as the $p_i^p$ (positive) and $(N-1)$ photo images of different subjects as the $p_i^n$ (negative). To address the degradation problem, we insert shortcut connection in proposed method which is proved in [He *et al.*, 2016]. The input size of the network is $224 \times 224$ and the first convolution layer creates 64 outputs with filter size of $7 \times 7$ shown in Figure 1. We build four basic blocks on the basis of [He

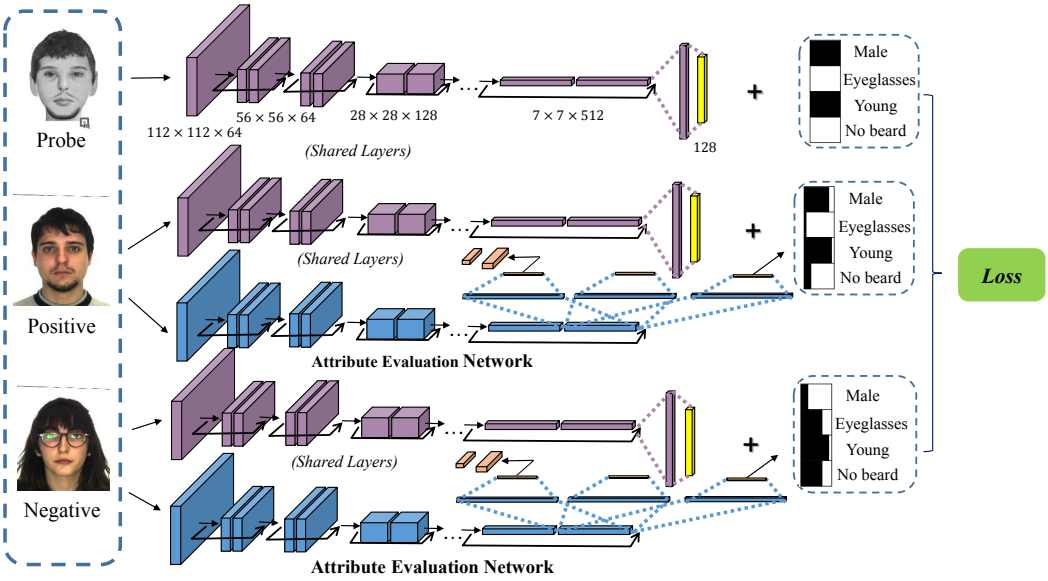


Figure 1: Overview of the proposed deep attribute guided heterogeneous face recognition.

*et al.*, 2016], where four kinds of convolutional layers separately generate 64, 128, 256 and 512 feature maps with filter size of $3 \times 3$ . To reduce the dimension of the mapping representations, we only employ one fully connected layers to generate features with 128 dimension. To get a better initial parameter, we pre-train the proposed network with the ImageNet database. Figure 1 shows the whole framework in the training stage. The three channels share the same parameters to make sure that heterogeneous face images could be mapped into the common space which effectively bridges the gap between different modalities.

In addition, attribute evaluation network is a modified ResNet network [2016] (18 convolutional layers, 6 Relu layers) with a batch normalization layer inserted after each convolutional layers. Each attribute evaluation network contains one specific fully connected layer, which is connected to the last fully connected layer as shown in figure 1. We pre-train this multi attributes evaluation network with CelebA dataset which contains 202,599 face images, each with 40 binary attributes labeled. With the same protocol in [Liu *et al.*, 2015], about $80\%$ of images are used to fine-tune, $10\%$ of images are used as validation data, and the rest of images are used as testing data. The performance of proposed attribute evaluation network for four selected attributes (Eyeglasses, Male, No beard and Young) is close to the state-of-art method [Han *et al.*, 2017].

In all experiments, we train the CNN using stochastic gradient descent and AdaGrad[Duchi *et al.*, 2011]. We start with a small learning rate of 0.0005 and reduce the learning rate by 0.1 every 5 epochs. We use a weight decay of 0.0005 and a momentum of 0.9.

## 3 Experiments

In this section, we evaluated the superior performance of the proposed method on two HFR scenarios tasks (composite sketch, resident ID card photo). These two kinds of heterogeneous face images both have inherent attributes information in the generation procedure without extra labor.

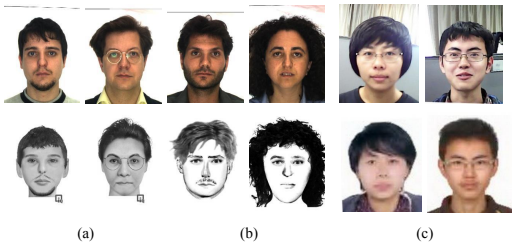


Figure 2: The illustration of heterogeneous face databases. (a): E-PRIP Composite Sketch database. (b): PRIP-VSGC Composite Sketch database. (c): NJU-ID Resident ID Card database.

We firstly investigate the effect of different parameters on the recognition performance. Then we evaluate the effectiveness of the proposed attribute guided triplet loss and detect the robustness of face attributes evaluation. Finally, we confirmed that our proposed approach achieved superior performance compared with state-of-art method on E-PRIP composite sketch database, PRIP-VSGC composite sketch database, NJU-ID resident ID card database.

### 3.1 Databases

In this section we would show two different heterogeneous scenarios. Example face images are shown in figure 2. For all experiment dataset, we randomly split the dataset into the training set and the testing set. The accuracies shown in this section are statistical results over 10 random partitions.

Extended PRIP Database (E-PRIP) contains 123 subjects, with photos from the AR database [Wang and Tang, 2009] and composite sketches created by FACES software. PRIP

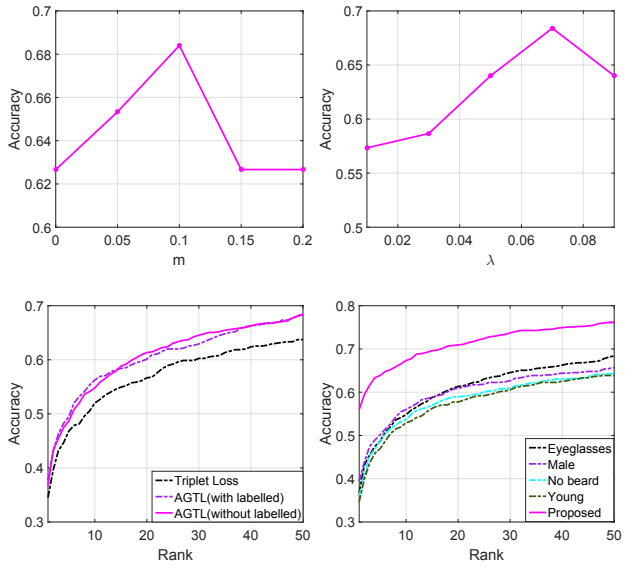


Figure 3: Left top subfigure shows the accuracies of different numbers of the parameter margin $m$ at rank-50; right top subfigure shows the accuracies of different numbers of the parameter $\lambda$ at rank-50; left bottom subfigure shows influence of the attributes evaluation performance; right bottom subfigure shows the accuracies by fusion different face attribues at rank-50. All the four experiments are conducted on the E-PRIP database with the 10,000 enlarged gallery.

Viewed Software-Generated Composite Database (PRIP-VSGC) also contains 123 subjects, with photos from the AR dataset and composite sketches created by Identi-Kit software. On these two kinds of face composite sketch databases, we follow the same protocol [Mittal *et al.*, 2015]. the 48 composite sketch-photo pairs from E-PRIP and PRIP-VSGC database are randomly selected as the training set, the rest pairs are the testing set.

NJU-ID database [Huo *et al.*, 2017] contains 256 persons. For each person, there are one card image with low resolution and one face photo from a high resolution digital camera. To evaluate perform on this database, we randomly select 100 pairs of ID card images and photo images as the training set, the rest 156 pairs are the testing set.

To make results much closer to the real scenarios, we evaluate the proposed method on the enlarged gallery [Peng *et al.*, 2017]. The enlarged gallery contains 10,000 face photo images of 5,329 subjects which mimic the real-world face retrieval scenarios. We follow the same strategy in [Peng *et al.*, 2017] and evaluation the proposed methods with enlarged gallery.

## 3.2 Experimental Settings

The parameters appeared in this paper are set as follows. All face images in the experiment are aligned according the eye centers. The proposed deep attribute guided HFR method related experiments conducted on Titan X GPU. And the other experiments are conducted on an Intel Core i7-4790 3.60GHz PC under MATLAB R2014a environment.

**Influence of parameter margin** $m$ We evaluate the effect of the margin $m$ in the E-PRIP composite dataset with enlarged gallery. In the left top subfigure of Figure 3, the effect of margin $m$ from a set of $\{0, 0.05, 0.1, 0.15, 0.2\}$ is illustrated when we set parameter $\lambda$ as 0.07. The recognition accuracy at rank 50 varies with different $m$. we find that when the margin $m$ reaches approximately 0.1 leading better performance.

**Influence of parameter** $\lambda$ We also investigated the effect of parameter $\lambda$ in the E-PRIP composite dataset with enlarged gallery. In the right top subfigure of Figure 3, the effect of parameter $\lambda$ from a set of 0.01, 0.03, 0.05, 0.07, 0.09 is explored when the margin $m$ is set at fixed value 0.1. In our proposed method, the parameter $\lambda$ balances the importance of the according attribute discriminability influence, which is important for recognition performance. We decide to set the hyper-parameter $\lambda$ as 0.07 in the following experiments. Thus we all set the parameter margin $m$ as 0.1 and set the hyper-parameter $\lambda$ as 0.07 in the formula of proposed Attribute Guided Triplet Loss (AGTL).

**Influence of the attributes evaluation performance** To prove the proposed DAG-HFR algorithm could effectively eliminate the adverse effect of evaluation error, we design an experiment and show the cumulative match curves in the left bottom subfigure of Figure 3. In this experiment, we label the training face photos attributes manually without evaluation error. And the attribute Eyeglasses prediction of this training face photos (123 images in AR dataset) is $90.24\%$ . The results show the performance of our proposed method is close to the performance of compared method with manual labeled, which shows that the proposed AGTL method is robust to the attributes evaluation. In addition, the accuracy of proposed AGTL outperforms original triplet loss, which means the proposed method could make representation more discriminative. In addition, an illustrate example of the influence face attributes evaluation is presented in Figure 5. From the ranked list acquired from different target functions, we could find it is indeed the attribute Eyeglasses results in a better matching list.

**Discussions on the fusion of different attributes** In this section, we choose four discriminative binary attributes (Eyeglasses, Male, No beard, Young) to explore the effectiveness of different attributes fusion method. The cumulative match score comparison of proposed attribute guided triplet loss with four different attributes are shown in the right bottom subfigure of Figure 3.Here we utilize the score fusion method to fuse four different attributes guided representation to achieve the best matching performance. It is reasonable to believe that the recognition performance could be further improved with more discriminative face attributes. Furthermore, we public an attribute annotations dataset for evaluation and to promote related researches.

## 3.3 Results on Multiple Databases

**Results on E-PRIP and PRIP-VSGC Composite Sketch Database** We compare the proposed approach with the state-of-art methods in Table 1 with protocol in [Mittal *et al.*, 2015]. It can be seen that the proposed method outperforms existing methods and reached rank-10 accuracy of $91.73\%$

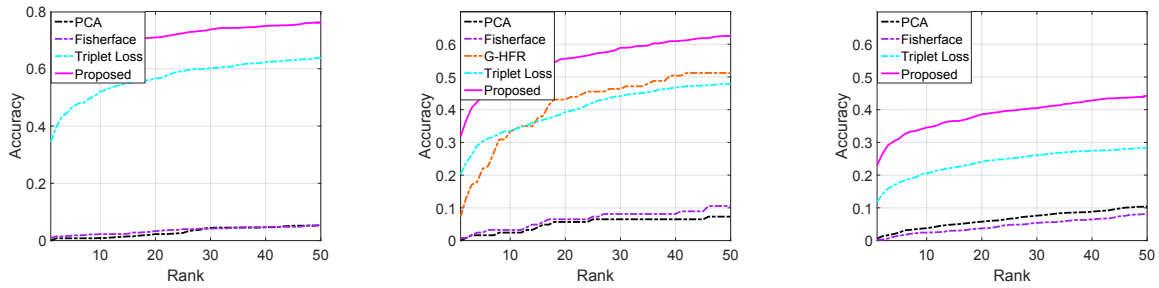


Figure 4: Left subfigure shows the cumulative match score comparison on E-PRIP database with enlarged gallery; middle subfigure shows the cumulative match score comparison on PRIP-VSGC database with enlarged gallery; right subfigure shows the cumulative match score comparison on NJU-ID database with enlarged gallery.

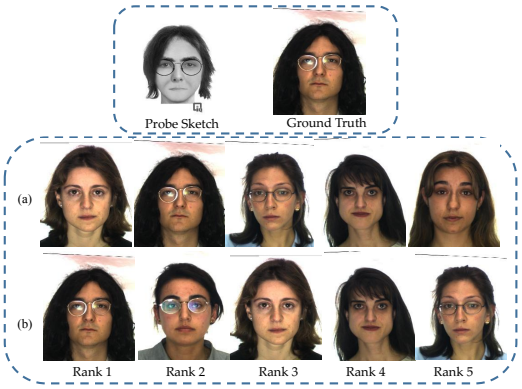


Figure 5: An illustration of the attribute guided representation on matching performance on E-PRIP database. (a):ranked list with triplet loss (b):ranked list with proposed AGTL (Eyeglasses)

| Algorithms | Accuracy (E-PRIP) | Accuracy (PRIP-VSGC) |
|---|---|---|
| Fisherface [Belhumeur *et al.*, 1997] | 35.30% | 21.87% |
| MCWLD [Bhatt *et al.*, 2012] | 24.00 % | 15.40 % |
| SSD-based [Mittal *et al.*, 2014] | 53.30% | 45.30% |
| Transfer Learning [Mittal *et al.*, 2015] | 60.20% | 52.00% |
| CNNs [Saxena and Verbeek, 2016] | 65.60% | 51.50% |
| **DAG-HFR** | **91.73%** | **86.27%** |

Table 1: Rank-10 recognition accuracies of the state-of-the-art approaches and our method on the E-PRIP and PRIP-VSGC databases without enlarged gallery.

and $86.27\%$ separately on the E-PRIP database and PRIP-VSGC database. In order to mimic the real-world face retrieval scenarios, we also evaluate methods performances with the enlarged gallery [Peng *et al.*, 2017]. On the E-PRIP database with enlarged gallery (shown in the left subfigure of Figure 4 , the proposed method achieves $76.13\%$ at rank-50, which is superior to baseline methods (Fisherface[1997], PCA) and original triplet loss. On the PRIP-VSGC database with enlarged gallery (shown in the middle subfigure of Figure 4), the proposed method achieves $62.53\%$ at rank-50, which outperform state-of-the-art method (G-HFR[2017]) of at least $14\%$.The reason is the proposed method could effectively extract more discriminative features by integrating face attributes information.

**Results on NJU-ID Resident ID Cards Database**

Considering the real-world scenarios, we directly match the resident ID card photos in the NJU-ID database with enlarged gallery (shown in the right subfigure of Figure 4). Due to the characters of this database, we choose the binary attribute male to enhance the recognition performance. Benefiting from the discriminative attributes information, the proposed method achieves $44.23\%$ at rank-50 which is superior to the triplet loss.

## 4 Conclusion

A deep attribute guided representation is proposed for heterogeneous face recognition in this paper. The proposed method could directly integrate the attribute discriminative information without manual face photos attributes labeling, which is suitable for large scale matching. Considering the adverse effect of incorrect attribute evaluation, we propose a attribute guided triplet loss to eliminate the negative influence. Experiments on E-PRIP database, PRIP-VSGC database and NJU-ID database illustrate the effectiveness of proposed method. The key benefit of the proposed DAG-HFR method is that the discriminative attributes information are crucial for HFR. In the future, we would evaluate the matching performance on more heterogeneous face recognition scenarios and integrate more discriminative facial attributes.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (under Grants 61501339, 61772402, 61671339, 61432014, U1605252), in part by Young Elite Scientists Sponsorship Program by CAST (under Grant 2016QNRC001), in part by Natural Science Basic Research Plan in Shaanxi Province of China (under Grant 2017JM6085), in part by Young Talent fund of University Association for Science and Technology in Shaanxi, China, in part by CCF-Tencent Open Fund (under Grant IAGR 20170103), in part by the Leading Talent of Technological Innovation of Ten-Thousands Talents Program under Grant CS31117200001, in part by the Fundamental Research Funds for the Central Universities under Grant XJS17086

## References


[Belhumeur *et al.*, 1997] 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.*, 19(7):711–720, 1997.

[Bhatt *et al.*, 2012] Himanshu S Bhatt, Samarth Bharadwaj, Richa Singh, and Mayank Vatsa. Memetically optimized mcwld for matching sketches with digital face images. *IEEE Transactions on Information Forensics and Security*, 7(5):1522–1535, 2012.

[Duchi *et al.*, 2011] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[Han *et al.*, 2017] Hu Han, Anil K Jain, Shiguang Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hu *et al.*, 2017] Guosheng Hu, Yang Hua, Yang Yuan, Zhihong Zhang, Zheng Lu, Sankha S Mukherjee, Timothy M Hospedales, Neil M Robertson, and Yongxin Yang. Attribute-enhanced face recognition with neural tensor fusion networks. In *Proc. Int. Conf. Comput. Vis.(ICCV)*, 2017.

[Huo *et al.*, 2017] Jing Huo, Yang Gao, Yinghuan Shi, Wanqi Yang, and Hujun Yin. Heterogeneous face recognition by margin-based cross-modality metric learning. *IEEE transactions on cybernetics*, 2017.

[Kan *et al.*, 2016] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):188–194, 2016.

[Klare *et al.*, Mar 2011] B. Klare, Z. Li, and A. Jain. Matching forensic sketches to mug shot photos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(3):639–646, Mar. 2011.

[Kumar *et al.*, 2011] Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

[Li *et al.*, 2015] Yan Li, Ruiping Wang, Haomiao Liu, Huajie Jiang, Shiguang Shan, and Xilin Chen. Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3819–3827, 2015.

[Lin and Tang, 2006] Dahua Lin and Xiaoou Tang. Inter-modality face recognition. In *European Conference on Computer Vision*, pages 13–26. Springer, 2006.

[Liu *et al.*, 2005] Qingshan Liu, Xiaoou Tang, Hongliang Jin, Hanqing Lu, and Songde Ma. A nonlinear approach for face sketch synthesis and recognition. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, pages 1005–1010 vol. 1, 2005.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[Lu *et al.*, 2017] Jiwen Lu, Venice Erin Liong, and Jie Zhou. Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[Mittal *et al.*, 2014] P. Mittal, A. Jain, G. Goswami, R. Singh, and M. Vatsa. Recognizing composite sketches with digital face images via ssd dictionary. In *IEEE International Joint Conference on Biometrics*, pages 1–6, 2014.

[Mittal *et al.*, 2015] P. Mittal, M. Vatsa, and R Singh. Composite sketch recognition via deep network. In *Proc. Int. Conf. Biom*, pages 1091–1097, 2015.

[Mittal *et al.*, 2017] Paritosh Mittal, Aishwarya Jain, Gaurav Goswami, Mayank Vatsa, and Richa Singh. Composite sketch recognition using saliency and attribute feedback. *Information Fusion*, 33:86–99, 2017.

[Ouyang *et al.*, 2014] Shuxin Ouyang, Timothy Hospedales, Yi-Zhe Song, and Xueming Li. Cross-modal face matching: beyond viewed sketches. In *Asian Conference on Computer Vision*, pages 210–225. Springer, 2014.

[Ouyang *et al.*, 2016] Shuxin Ouyang, Timothy M Hospedales, Yi-Zhe Song, and Xueming Li. Forgetmenot: memory-aware forensic facial sketch matching. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5571–5579, 2016.

[Peng *et al.*, 2017] Chunlei Peng, Xinbo Gao, Nannan Wang, and Jie Li. Graphical representation for heterogeneous face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(2):301–312, 2017.

[Saxena and Verbeek, 2016] Shreyas Saxena and Jakob Verbeek. Heterogeneous face recognition with cnns. In *European Conference on Computer Vision*, pages 483–491. Springer, 2016.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[Tang and Wang, 2003] Xiaoou Tang and Xiaogang Wang. Face sketch synthesis and recognition. In *Computer vision, 2003. proceedings. ninth ieee international conference on*, pages 687–694. IEEE, 2003.

[Wang and Tang, 2009] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(11):1955–1967, 2009.