2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM) | 978-1-6654-2678-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/IMCOM53663.2022.9721719

Face Photo-Sketch Recognition Using Bidirectional Collaborative Synthesis Network

Seho Bae , Nizam Ud Din , Hyunkyu Park , and Juneho Yi

Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Republic of Korea
{*bseho, nizam, mjss016, jhyi*}*@skku.edu*

***Abstract*—This research features a deep learning based framework to address the problem of matching a given face sketch image against a face photo database. The photo-sketch matching problem is challenging because 1) modality gap between photo and sketch is very large, and 2) the number of paired photo/ sketch data is insufficient to train deep network. To circumvent the problem of large modality gap, our approach is to use an intermediate latent space between the two modalities. We effectively align the distributions of the two modalities in this latent space by employing a bidirectional (photo $\rightarrow$ sketch and sketch $\rightarrow$ photo) collaborative synthesis network. A StyleGAN-like architecture is utilized to make the intermediate latent space be equipped with rich representation power. To resolve the problem of insufficient training samples, we introduce a three-step training scheme. Extensive evaluation on public composite face sketch database confirms superior performance of our method compared to existing state-of-the-art methods. The proposed methodology can be employed in matching other modality pairs.**

***Index Terms*—Face photo-sketch recognition, Face photo-sketch synthesis, GAN**

I. Introduction

The goal of this work is to find the best matching photos for a given sketch in a face database, especially for software generated composite sketches. An important application of such systems is to assist law-enforcement agencies. In many cases, face photo of a suspect is unavailable for criminal investigation. Instead, the only clue to identify suspect is software generated composite sketch or hand-drawn forensic sketch based on the description by an eye-witness. Therefore, an automatic method which retrieves the best matching photos from face database for a given sketch is necessary to quickly and accurately identify a suspect.

Successful photo-sketch matching depends on the solution to how to effectively deal with large modality gap between photo and sketch modalities.Moreover, insufficiency of sketch samples for training makes photo-sketch recognition an extremely challenging task.

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2020R1F1A1048438). This research was results of a study on the "HPC Support" Project, supported by the 'Ministry of Science and ICT' and NIPA.

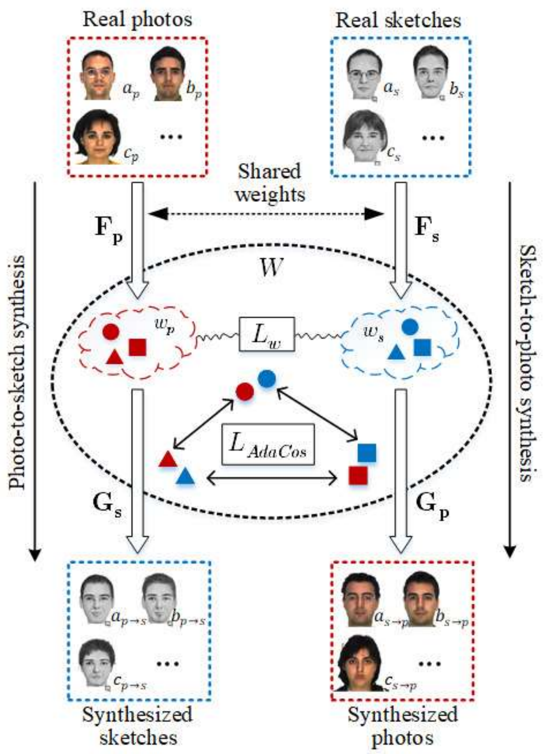

Fig. 1. Our proposed framework takes advantage of a bidirectional photo/sketch synthesis network to set up an intermediate latent space as an effective homogeneous space for face photo-sketch recognition. We employ a StyleGAN-like architecture to make the intermediate latent space be equipped with rich representational power. The mapping networks, $\mathbf{F}_p$ and $\mathbf{F}_s$, learn to encode photo and sketch images into their respective intermediate latent codes, w_p and w_s. We learn AdaCos [1] to enforce the separability of latent codes of different identity in the angular space for the photo-sketch recognition task.

As to classical photo-sketch recognition, generative approaches [2–4] bring both modalities into a single modality by transforming one of the modalities to the other (either photo to sketch or vice versa) before matching. The main drawback of these methods is their dependency on the quality of the synthetic output, which most of the time suffers due to large modality gap between photos and sketches. On the other hand, discriminative approaches attempt to extract modality-invariant features, or learn a common subspace where both photo and sketch modalities are aligned [5–13]. Although these

978-1-6654-2678-7/22/$31.00 ©2022 IEEE

methods formulate photo-sketch recognition through modality invariant features or a common subspace, their performances are not satisfactory because 1) the distributions of the two modalities are not well aligned in the common feature space and 2) their feature vectors or common spaces fail to provide rich representation capacity. Recent deep learning based face photo-sketch recognition methods [7, 10, 14–20] perform well compared to classical approaches. However, employing deep learning techniques for face photo-sketch recognition is very challenging because of insufficient training data.

Recently, Col-cGAN [21] proposed a bidirectional face photo-sketch synthesis network. They generate synthetic outputs by using a middle latent domain between photo and sketch modalities. However, their middle latent domain does not provide enough representational power of both modalities. On the other hand, StyleGAN [22] produces extremely realistic images by proposing a novel generator architecture. Instead of feeding the input latent code z directly into the generator, the StyleGAN network first transforms it into an intermediate latent space, W, via a mapping network. This disentangled intermediate latent space, W, offers the StyleGAN generator more control and representational capabilities. Noting the strong representation power of the latent code space of StyleGAN, we opt to use a StyleGAN-like bidirectional architecture for setting up an intermediate latent space for our photo-sketch recognition problem.

In this paper, we propose a novel method that exploits an intermediate latent space, W, between the photo and sketch modalities as shown in Figure 1. We employ a bidirectional collaborative synthesis network of the two modalities to set up the intermediate latent space where the distributions of the two modalities are effectively aligned. Also, the StyleGAN-like architecture we utilize enables the intermediate latent space to have strong representational power to successfully match the two modalities.

In Figure 1, the mapping networks, $\mathbf{F}_p$ and $\mathbf{F}_s$, learn the intermediate latent codes $w_p, w_s \in W$. To form a homogeneous intermediate space, W, we constrain the intermediate features more symmetrical, using ℓ_1 distance between the intermediate latent codes of photo and sketch The intermediate latent space also makes use of feedback from the style generators that translate photo-to-sketch/sketch-to-photo. Hereby enabling the intermediate latent space to have rich representational capacity for both photo and sketch. Once this intermediate latent space is successfully set up, we can then directly take advantage of any state-of-the-art face recognition methods. In our case, we employ AdaCos loss [1].

Moreover, we use a three-step training scheme to resolve the problem of very limited number of training sketch samples. In the first step, we only learn image-to-image translation without AdaCos on paired photo-sketch samples. This serves the purpose of learning an initial intermediate latent space. Then, in the second step, we pre-train the photo mapping network, $\mathbf{F}_p$, only with AdaCos, using a publicly available large photo dataset. This helps our model overcoming the problem of insufficient sketch samples to train our deep network robustly for the target task. Lastly, we fine tune the full network on a target photo/sketch dataset. More details of the model training are discussed in section III-B.

The main contributions of our work are summarized as follows.

- We propose a novel method for photo-sketch matching that exploits an intermediate latent space between the photo and sketch modalities:
  - The intermediate latent space is built through a bidirectional collaborative synthesis network.
  - This latent space has rich representational power for photo/sketch recognition due to a StyleGAN-like architecture.
- A three-step training scheme helps overcoming the problem of insufficient sketch training samples.
- Extensive evaluation on challenging publicly available composite face sketch databases shows superior performance of our method compared with state-of-art methods.

The rest of the paper is organized as follows. Section II states related works. We elaborate details of our method in section III. Section IV shows experimental results.

II. Related work

The face photo-sketch recognition problem has been extensively studied in recent years. Researchers have studied sketch recognition for various face sketch categories such as hand-drawn viewed sketch, hand-drawn semi-forensic sketch, hand-drawn forensic sketch, and software generated composite sketch. Compared to hand-drawn viewed sketches, other sketch categories have much larger modality gap due to the errors that come from forgetting (semi-forensic/forensic), understanding of description (forensic), or limitation of components in software (composite). Recent researches focus on more challenging composite and forensic sketches.

Trivial recognition methods can be categorized into generative and discriminative approches.

Generative methods convert images from one modality into the other modality, usually from sketch to photo, before matching. Then, a simple homogeneous face recognition method can be used for matching. Various techniques have been utilized for synthesis such as Markov random field model [2], local linear embeding (LLE) [3], and multi-task gaussian process regression [4]. However, these approaches recognition performance is strongly reliant on their synthesis results, which most of the time suffers due to the large modality gap between the two modalities

Discriminative methods attempt to learn a common subspace or extract particular features in order to reduce the modality gap between photos and sketches of same identity while increasing the gap between different identities. Representative methods in this category include partial linear square(PLS) [5, 6], coupled information-theoretic projection (CITP) [7], local feature-based discriminant analysis (LFDA) [8], canonical correlation analysis (CCA) [9], and self similarity descriptor (SSD) dictionary [10]. Han *et al.* [11]

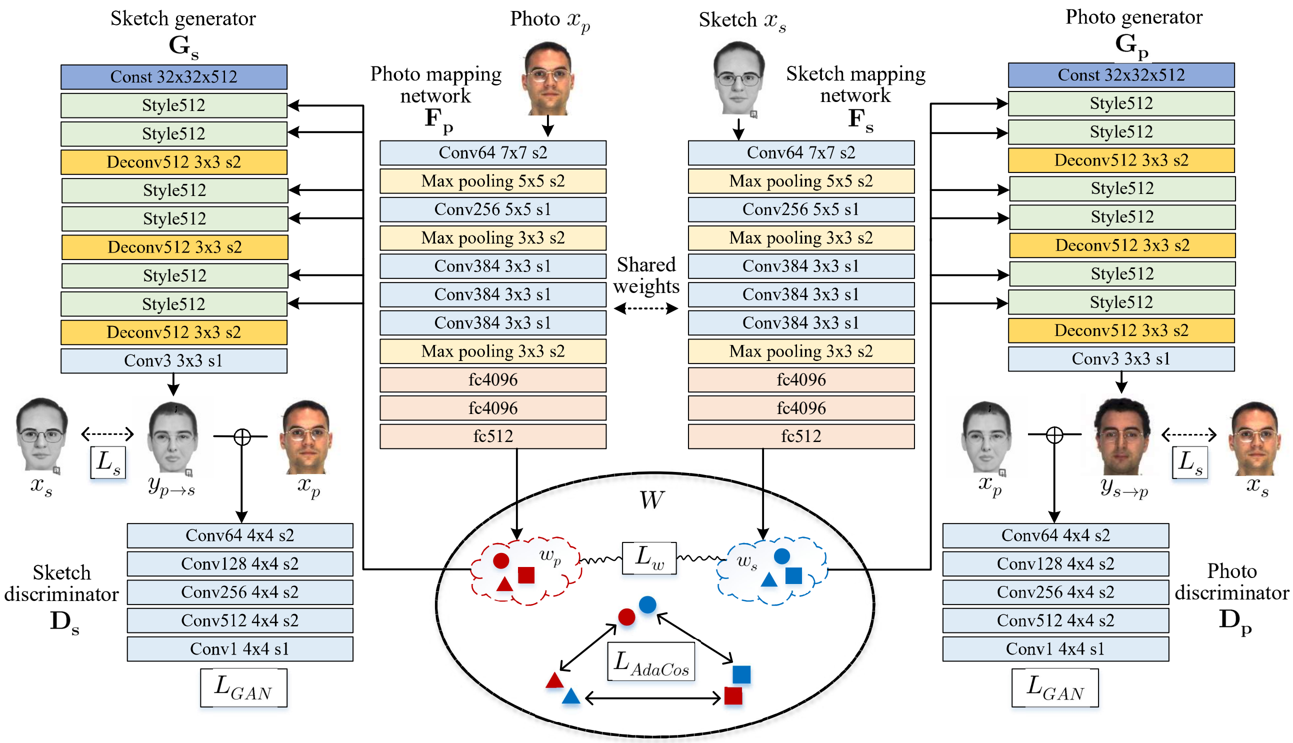

Fig. 2. The overall architecture of the proposed network. Mapping networks, $\mathbf{F}_p$ and $\mathbf{F}_s$, map photo and sketch images to intermediate latent codes w_p and w_s. These latent codes are then fed into the two opposite style generators $\mathbf{G}_s$ and $\mathbf{G}_p$. $\mathbf{G}_s$ generates sketch from photo, $y_{p\to s}$, while $\mathbf{G}_s$ generates photo from sketch, $y_{s\to p}$. The collaborative loss L_w, which is ℓ_1 distance between w_p and w_s of same identity, constrains the intermediate modality features more symmetrical. Through this strategy, we learn an intermediate latent space, W, that retain the common and representational information of photo and sketch. We apply AdaCos loss, L_{AdaCos}, to the intermediate latent space, W, directly to perform photo-sketch recognition by comparing the cosine distance between intermediate latent features, w_p and w_s.

compute the similarity between a photo and composite sketch using a component-based representation technique. Multi-scale circular Weber's local descriptor (MCWLD) is utilized in Bhatt *et al.* [12] to solve semi-forensic and forensic sketch recognition problem. In graphical representation based heterogeneous face recognition (G-HFR) [13], the authors graphically represented heterogeneous image patches by employing Markov networks, and designed a similarity metric for matching. These methods fail when the learned feature/common subspace could not have enough representational capacity for both photo and sketch modalities. In contrast, our method projects photo and sketch on homogeneous intermediate space where the distribution of the two modalities better aligned with rich representational power.

Over the past few years, deep learning based algorithms have been developed for face photo-sketch recognition [7, 10, 14–19]. Kazemi *et al.* [14] and Iranmanesh *et al.* [15] proposed attribute-guided approaches by introducing attribute-centered loss function and joint loss function of identity and facial attribute classification, respectively. Liu *et al.* designed an end-to-end recognition network using coupled attribute guided triplet loss (CAGTL). It plausibly eliminates defects of incorrectly estimated attributes [16] during training. Iterative local re-ranking with attribute guided synthesis based on GAN is introduced in [17]. Peng *et al.* proposed DLFace [18] which is a local descriptor approach based on deep metric learning while in [19], a hybrid feature model was employed by fusing traditional HOG feature with deep feature. The largest obstacle to utilizing deep learning techniques for face photo-sketch recognition is scarcity of sketch data. Even the largest public viewed sketch database [7] has only 1,194 pairs of sketch and photo, and the composite sketch database [10] has photos and sketches of 123 identities. To overcome this problem, most approaches employ relatively shallow network, data augmentation, or pre-training on a large-scale face photo database.

Recently, cosine-based softmax losses [1, 23–25] have achieved great success in face photo recognition. SphereFace [23] applied multiplicative penalty to the angles between the deep features and their corresponding weights. Follow-up studies improved the performance by changing the penalising measure to additive margin in cosine [24] and angle [25]. AdaCos [1] outperforms previous cosine-based softmax losses by leveraging an adaptive scale parameter to automatically strengthen the supervision during training. However, direct application of these methods to photo-sketch recognition is not satisfactory because they have not properly dealt with the modality gap.

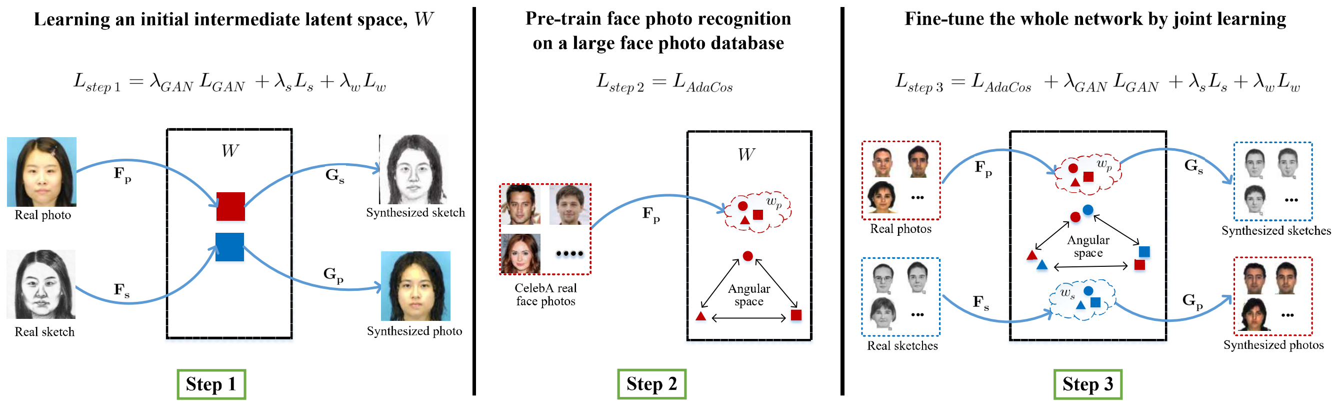

Fig. 3. A three-step training scheme to overcome the problem of insufficient amount of paired photo-sketch training samples. We employ three-step training. Step 1: Pre-train the bidirectional photo/sketch synthesis network to learn an initial intermediate latent space, W, between photo and sketch. Step 2: Pre-train the photo mapping network, $\mathbf{F}_p$, on a large face photo database. Step 3: Fine-tune the whole network on the target photo/sketch database.

III. Proposed Method

Our proposed framework takes advantage of a bidirectional photo/sketch synthesis network to set up an intermediate latent space as an effective homogeneous space for face photo-sketch recognition. Mutual interaction of the two opposite synthesis mappings occurs in the bidirectional collaborative synthesis network. The complete structure of our network is illustrated in Figure 2. Our network consists of mapping networks $\mathbf{F}_p$ and $\mathbf{F}_s$, style generators $\mathbf{G}_p$ and $\mathbf{G}_s$, and discriminators $\mathbf{D}_p$ and $\mathbf{D}_s$. $\mathbf{F}_p$ and $\mathbf{F}_s$ share their weights.

The mapping networks, $\mathbf{F}_p$ and $\mathbf{F}_s$, learn to encode photo and sketch images into their respective intermediate latent codes, w_p and w_s. Then, w_p and w_s are fed into the two opposite style generators $\mathbf{G}_s$ and $\mathbf{G}_p$ to map photo-to-sketch and sketch-to-photo, respectively. We employ a StyleGAN-like architecture to make the intermediate latent space be equipped with rich representational power. We also introduce a loss function to regularize the intermediate latent codes of two modalities, enabling them to learn a same feature distribution. Through this strategy, we learn a homogeneous intermediate feature space, W, that shares common information of the two modalities, thus producing best results for heterogeneous face recognition. To enforce latent codes in W separable in the angular space, we learn AdaCos [1] for the photo-sketch recognition task.

$\mathbf{F}_p$ and $\mathbf{F}_s$ employ a simple encoder architecture that contains convolution, max pooling and fully connected layers. The style generators, $\mathbf{G}_p$ and $\mathbf{G}_s$, consist of several style blocks and deconvolution layers as in [22]. However, unlike [22], we do not use noise inputs and progressively growing architecture because the sole purpose of our style generators is to help the homogeneous intermediate latent space retain common representational information of the two modalities for reducing the modality gap between them. Our style generator architecture is very light as compared to that of StyleGAN due to limited number of training samples. The discriminators, $\mathbf{D}_p$ and $\mathbf{D}_s$, distinguish generated photo/sketch and real samples by taking corresponding concatenated photo and sketch. We use PatchGAN architecture [26] of 70x70. Unlike the discriminator in [21], our discriminator uses *Instance normalization* instead of *Batch normalization*.

A. Loss functions

The joint loss function used to train our framework is defined as:

$$L = L_{AdaCos} + \lambda_{GAN} L_{GAN} + \lambda_s L_s + \lambda_w L_w \tag{1}$$

GAN loss function, L_{GAN} [27], along with the similarity loss, L_s, are used to train the bidirectional photo/sketch synthesis part of the whole network. L_{GAN} helps generating real and natural-looking synthetic outputs while the similarity loss, L_s, measures pixel-wise ℓ_1 distance between generated and real photo/sketch images. To regularize and enforce the same distribution for photo, w_p, and sketch, w_s, in the intermediate latent space, we introduce a collaborative loss, L_w. It minimizes ℓ_1 distance between w_p and w_s of the same identity. We use AdaCos loss function [1], L_{AdaCos}, to learn identity recognition. It measures the angular distance in the W space. It is minimized for intra-class features and maximized for inter-class features.

λ_{GAN}, λ_s, and λ_w in Eq. (1) control the relative importance of each loss function in the bidirectional photo/sketch synthesis task. We used $\lambda_{GAN} = 1$, $\lambda_s = 10$, and $\lambda_w = 1$ in our experiments.

B. Training

To overcome the problem of insufficient amount of paired photo/sketch training data, we introduce a simple and effective three-step training scheme as shown in Figure 3. In step 1, we train the bidirectional photo/sketch synthesis network using paired photo-sketch training samples to set up an initial homogeneous intermediate latent space, W. We use our joint loss function in Eq. (1), excluding the AdaCos loss function, L_{AdaCos}. In step 2, we pre-train the photo mapping network, $\mathbf{F}_p$, using AdaCos loss only on the publicly available large

TABLE I
RANK 50 RECOGNITION ACCURACY (%) ON THE E-PRIP DATABASE WITH A GALLERY SIZE 1,500.

Method	Faces (In)	Identikit (As)
Kazemi *et al.* [14]	77.50	81.50
Iranmanesh *et al.* [15]	80.00	83.00
Ours	**93.86**	**90.40**

TABLE II
RANK 50 RECOGNITION ACCURACY (%) ON THE E-PRIP DATASET WITH A GALLERY SIZE 10,075.

Method	Faces (In)	Identikit (As)
G-HFR [13]	-	51.22
DLFace [18]	70.00	58.93
CAGTL [16]	78.13	67.20
Ours	**92.78**	**88.26**

photo database CelebA [28] to overcome the problem of insufficient sketch training samples. Then, we train our full network in step 3 using the whole joint loss function in Eq. (1) on target photo/sketch samples.

IV. EXPERIMENTS

A. Data description and implementation details

We have conducted our experiments using the e-PRIP composite sketch dataset. The e-PRIP [10] dataset consists of four different composite sketch sets of 123 identities. However, only two of them are publicly available: the composite sketches created by an Indian user adopting the FACES tool [29], and an Asian artist using the Identi-Kit tool [30]. We have used 48 identities for training and the remaining 75 identities for test.

We aligned all photos and sketches by eye position and initially cropped to 272x272. Then, they are randomly cropped to 256x256 during training. We optimize our network using Adam optimizer with the learning rate of 0.0002 and batch size 8, in step 1 and 3 of training. We use the learning rate 0.0005 and batch size 32 in step 2. We train our network for 3,000 epochs on the CUFS [31] viewed sketch database in step 1 of training, 50 epochs on CelebA [28] in step 2, and 3,000 epochs on the target database in step 3.

The recognition accuracies of our network presented in the following sections are average results over five experiments with random partitions.

B. Photo-sketch recognition results

In this section, we compare the performance of our method with that of representative state-of-the-art photo-sketch matching methods on the two subsets of e-PRIP dataset [10]. Let us denote them FACES (In) and Identikit (As), respectively. We perform the experiments with an extended gallery to a mimic real law-enforcement scenario where multiple numbers of suspects are selected from a large photo database. With extended gallery setting, rank 50 accuracy is most commonly used criteria. Thus we compared rank 50 accuracies. While some photos in extended galleries of previous works are not publicly available, we have tried to mimic their gallery as close as possible using publicly available databases for fair comparison.

Following [14] and [15], we have constructed an extended gallery of 1,500 subjects including probe images by using photos from ColorFERET [32], Multiple Encounter Dataset (MEDS) [33], and CUFS [31]. The results are presented in Table I where the accuracies for Kazemi *et al.* and Iranmanesh *et al.* are obtained from their CMC curves. Our method achieved 93.86% rank 50 accuracy on Faces (In) which was 13.86% higher than [15]. On Identikit (As), our method achieved 90.40% which outperformed SOTA.

To compare the performance with [13, 18] and [16], we have built another extended gallery of 10,000 subjects using face photos collected from the aforementioned photo databases and the labeled faces in the wild-a (LFW-a) database [34]. The test gallery set contains the total of 10,075 face photos. Table II shows the comparison results of our method with the previous state-of-the-art representative methods. As can be seen, our method shows the far better performance of 92.78% and 88.26% rank 50 accuracies on Faces (In) and Identikit (As), respectively, with large margins. These results show that our bidirectional collaborative StyleGAN-like Synthesis Network learns an effective intermediate latent space with rich representational power for face photo-sketch recognition task.

C. Effect of bidirectional collaborative synthesis of photo-to-sketch and sketch-to-photo

To verify the effectiveness of our StyleGAN-like bidirectional collaborative synthesis network on the recognition task, we give comparison with three different versions from the full network. In the first version, we removed the style generators, $\mathbf{G}_s$ and $\mathbf{G}_p$, from the network in Figure 2 and train the mapping networks, $\mathbf{F}_p$ and $\mathbf{F}_s$, using AdaCos loss function. That is, the first version could not take any advantage of synthesis network. For this version, the mapping networks are pre-trained for 50 epochs on the CelebA photo database [28], then fine-tuned for 3,000 more epochs on the target database. For the second and third versions, we trained a unidirectional synthesis based photo-sketch recognition network by using only one of the style generators, either $\mathbf{G}_s$ or $\mathbf{G}_p$. These two versions employed the three-step training scheme as in the full network.

The comparison results in Table III indicate that the addition of either photo or sketch generator improves the recognition accuracy. The unidirectional sketch-to-photo network shows better performance than the unidirectional photo-to-sketch network. This is because sketch-to-photo network translates the information-poor input to information-rich output, thus providing better representational feedback to the intermediate latent space as compared to photo-to-sketch network. However, it still cannot provide enough representational power.

Our full network which exploited the bidirectional collaborative synthesis network dramatically improved the recognition performance. It is because our bidirectional synthesis network warrants the intermediate latent space to have important rep-

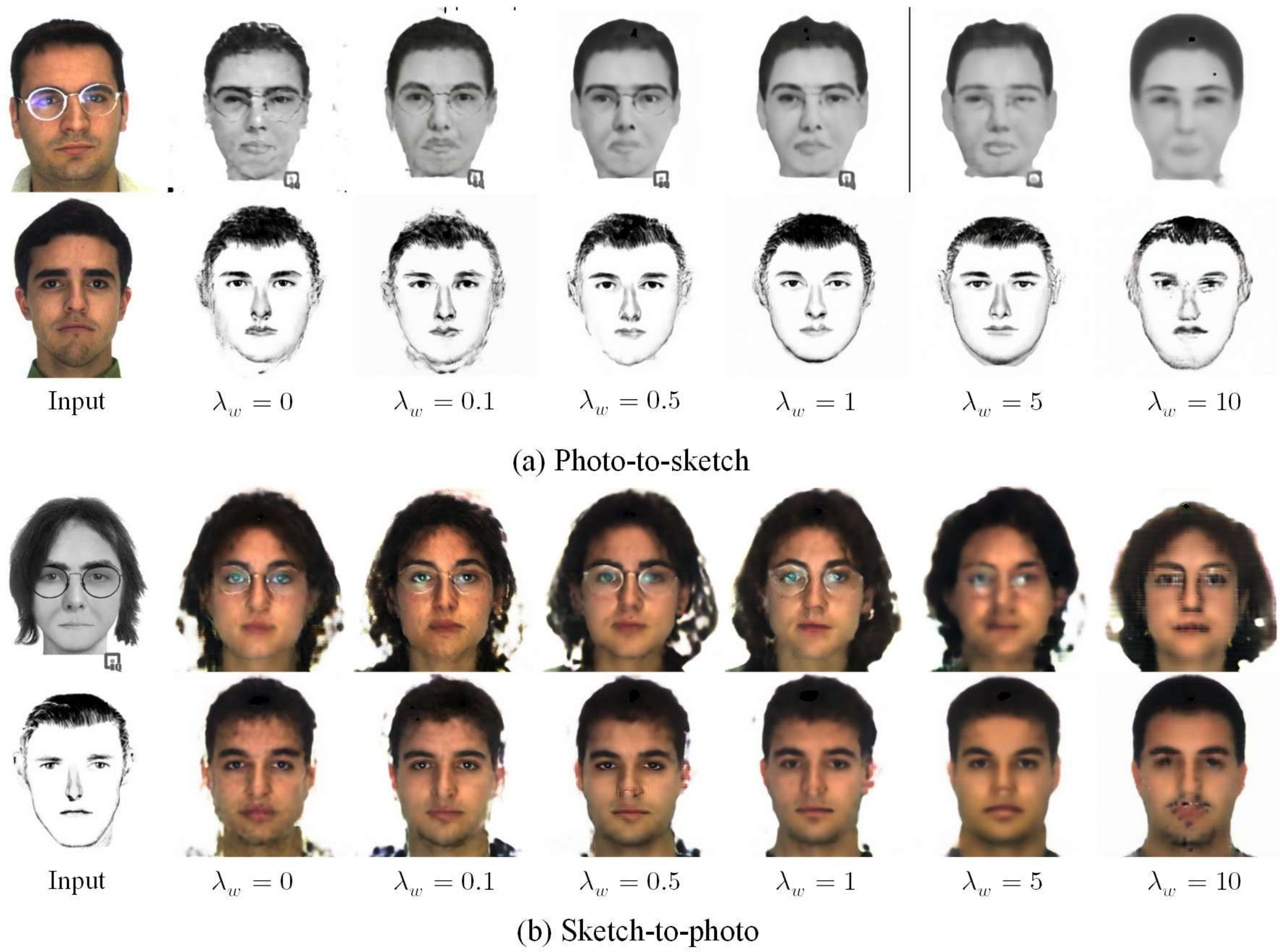

Fig. 4. Synthesis results of our style generators for different values of λ_w. (a) photo-to-sketch synthesis and (b) sketch-to-photo synthesis. First and second rows are for Faces (In), while third row is for Identikit (As). Images collapse with high λ_w so that network could not learn representational information of photo and sketch. λ_w=1 shows the best synthesis results. (Please view in color.)

TABLE III
RANK 50 RECOGNITION ACCURACY (%) ON THE E-PRIP DATASET WITH A GALLERY SIZE 1,500 FOR THE SYNTHESIS NETWORK.

Method	Faces (In)	Identikit (As)
Only mapping networks	19.74	43.72
Photo-to-sketch (with $\mathbf{G}_p$ removed)	68.54	61.58
Sketch-to-photo (with $\mathbf{G}_s$ removed)	73.84	73.88
Our full network (with both $\mathbf{G}_p$ and $\mathbf{G}_s$)	**93.86**	**90.40**

resentational information by utilizing the mutual interaction between the two opposite mappings.

D. Effect of three-step training scheme

To validate the effectiveness of the proposed three-step training scheme, we compare three different training settings in Table IV. For this, we train our model 1) using only step 3, that is, without pre-training, 2) using step 2 and step 3, and 3) using all the three steps. We can see that there is significant improvement in recognition accuracy when using pre-training (step 2), especially for Faces (In) dataset. This shows the power of large-scale pre-training in solving data scarcity problem. The combination of all the three training steps further boosts the recognition performance. Step 1 provides an effective initialization of the intermediate latent space between photo and sketch for large-scale training in step 2. As the last row in Table IV shows, our three-step training strategy effectively overcomes the problem of insufficient sketch training samples.

E. Collaborative loss, L_w

In this section, we analyze the effect of collaborative loss, L_w, on the recognition accuracy. We experimented our network as we change the value of λ_w. Table V shows the results for different values of λ_w on the extended gallery setting of 1,500 samples.

The performance for $\lambda_w = 0$ is poor. $\lambda_w = 0$ means that our network is not using collaborative loss L_w. The network is unable to constrain the two mappings symmetrical. The accuracy improves when we increase the value of λ_w as can be seen in Table V. Through many experiments, we

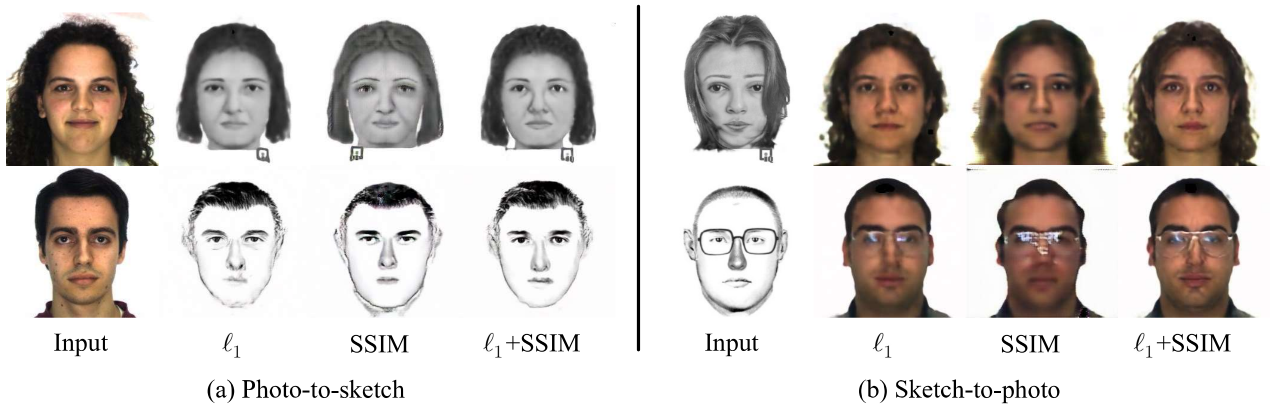

Fig. 5. Synthesis results of our style generators for three different versions of L_s. (a) photo-to-sketch synthesis and (b) sketch-to-photo synthesis. First and second rows are trained on Faces (In), while third and fourth rows are trained on Identikit (As). (Please view in color.)

TABLE IV
RANK 50 RECOGNITION ACCURACY (%) ON THE E-PRIP DATASET WITH A GALLERY SIZE 1,500 FOR TRAINING SCHEME.

Method	Faces (In)	Identikit (As)
Without pre-training (step 3 only)	25.32	46.14
Two-step training (step 2 + step 3)	90.66	89.60
Three-step training (step 1 + step 2 + step 3)	**93.86**	**90.40**

TABLE V
RANK 50 RECOGNITION ACCURACY (%) ON THE E-PRIP DATASET WITH A GALLERY SIZE 1,500 FOR λ_w.

Method	Faces (In)	Identikit (As)
λ_w = 0	72.00	66.40
λ_w = 0.1	89.32	82.68
λ_w = 0.5	89.60	85.60
λ_w = 1	**93.86**	**90.40**
λ_w = 5	85.34	84.28
λ_w = 10	83.72	83.74

have found that $\lambda_w = 1$ produces the best result for our task. These results show that our collaborative loss helps regularizing the intermediate latent representations of the two different modalities, effectively aligning the two modalities in the intermediate latent space. However, as λ_w gets too large, the performance degrades as can be seen in Table V. We think that a large value of λ_w emphasizes too much on making latent codes symmetrical, and breaks the learning balance of the latent space between representational capacity and symmetrical mapping.

Figure 4 shows examples of synthesis results produced by our style generators for different values of λ_w. There is a general trend that better synthesis results yield better recognition accuracies. For $\lambda_w = 10$, the results collapsed to the same synthesis result for most of the target samples. This shows that too much weightage to the collaborative loss strongly enforces the same latent distribution while the representational capacity of the latent space relatively ignored.

F. Similarity loss, L_s

Figure 5 shows the results produced by our style generators for three simple variations of L_s. First, we used pixel-wise ℓ_1 distance only as our L_s. Second, we used only patch-wise structural similarity (SSIM) loss [35]. Third, we employed

TABLE VI
RANK 50 RECOGNITION ACCURACY (%) ON THE E-PRIP DATASET WITH A GALLERY SIZE 1,500 FOR L_s.

Method	Faces (In)	Identikit (As)
ℓ_1	**93.86**	**90.40**
SSIM	81.86	79.74
ℓ_1 + SSIM	91.98	89.34

SSIM loss along with ℓ_1 distance for L_s. Figure 5 shows that using only SSIM loss for L_s produces the worst synthetic results, yielding the lowest recognition accuracy as can be seen in Table VI. On the other hand, ℓ_1 produces the best recognition results compared to the other two settings. Our observation is that SSIM loss provides extra structural information for synthesis, but it does not help for recognition task. Thus, we opt to use only ℓ_1 distance as our L_s in the joint loss function in Eq. (1).

V. CONCLUSION

We proposed a novel deep learning based face photo-sketch recognition method by exploiting a homogeneous intermediate latent space between photo and sketch modalities. For this, we introduce a bidirectional photo/sketch synthesis network based on a StyleGAN-like architecture. In addition, we employ a simple three-step training scheme to overcome the problem of insufficient photo/sketch data. The experiment results have verified the effectiveness of our method, outperforming the representative state-of-the-arts. Our method shows great promise in matching pairs of other different modalities.

REFERENCES

[1] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, "Adacos: Adaptively scaling cosine logits for effectively learning deep face representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 823–10 832.

[2] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 11, pp. 1955–1967, 2008.

[3] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," in *2005 IEEE Computer Society conference on computer vision and pattern recognition (CVPR'05)*, IEEE, vol. 1, 2005, pp. 1005–1010.

[4] S. Ouyang, T. M. Hospedales, Y.-Z. Song, and X. Li, "Forgetmenot: Memory-aware forensic facial sketch matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5571–5579.

[5] A. Sharma and D. W. Jacobs, "Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch," in *CVPR 2011*, IEEE, 2011, pp. 593–600.

[6] J. Choi, A. Sharma, D. W. Jacobs, and L. S. Davis, "Data insufficiency in sketch versus photo face recognition," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 2012, pp. 1–8.

[7] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *CVPR 2011*, IEEE, 2011, pp. 513–520.

[8] B. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 3, pp. 639–646, 2010.

[9] S. Ouyang, T. Hospedales, Y.-Z. Song, and X. Li, "Cross-modal face matching: Beyond viewed sketches," in *Asian Conference on Computer Vision*, Springer, 2014, pp. 210–225.

[10] 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*, IEEE, 2014, pp. 1–6.

[11] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain, "Matching composite sketches to face photos: A component-based approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 191–204, 2012.

[12] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetically optimized mcwld for matching sketches with digital face images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1522–1535, 2012.

[13] C. Peng, X. Gao, N. Wang, and J. Li, "Graphical representation for heterogeneous face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 2, pp. 301–312, 2016.

[14] H. Kazemi, S. Soleymani, A. Dabouei, M. Iranmanesh, and N. M. Nasrabadi, "Attribute-centered loss for soft-biometrics guided face sketch-photo recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 499–507.

[15] S. M. Iranmanesh, H. Kazemi, S. Soleymani, A. Dabouei, and N. M. Nasrabadi, "Deep sketch-photo face recognition assisted by facial attributes," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, IEEE, 2018, pp. 1–10.

[16] D. Liu, X. Gao, N. Wang, J. Li, and C. Peng, "Coupled attribute learning for heterogeneous face recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[17] D. Liu, X. Gao, N. Wang, C. Peng, and J. Li, "Iterative local re-ranking with attribute guided synthesis for face sketch recognition," *Pattern Recognition*, vol. 109, p. 107 579,

[18] C. Peng, N. Wang, J. Li, and X. Gao, "Dlface: Deep local descriptor for cross-modality face recognition," *Pattern Recognition*, vol. 90, pp. 161–171, 2019.

[19] J. Xu, X. Xue, Y. Wu, and X. Mao, "Matching a composite sketch to a photographed face using fused hog and deep feature models," *The Visual Computer*, pp. 1–12, 2020.

[20] H. Cheraghi and H. J. Lee, "Sp-net: A novel framework to identify composite sketch," *IEEE Access*, vol. 7, pp. 131 749–131 757, 2019. DOI: 10.1109/ACCESS.2019.2921382.

[21] M. Zhu, J. Li, N. Wang, and X. Gao, "A deep collaborative framework for face photo–sketch synthesis," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 10, pp. 3096–3108, 2019.

[22] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[23] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[24] H. Wang *et al.*, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[27] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[28] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, Dec. 2015.

[29] *FACES 4.0, IQ Biometrix 2011 [Online]*, Available: http://www.iqbiometrix.com/.

[30] *Identi-Kit, Identi-Kit Solutions 2011 [Online]*, Available: http://www.identikit.net/.

[31] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Proceedings Ninth IEEE International Conference on Computer Vision*, IEEE, 2003, pp. 687–694.

[32] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[33] A. P. Founds, N. Orlans, W. Genevieve, and C. I. Watson, "Nist special databse 32-multiple encounter dataset ii (meds-ii)," Tech. Rep., 2011.

[34] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 10, pp. 1978–1990, 2010.

[35] J. Snell, K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer, and R. S. Zemel, "Learning to generate images with perceptual similarity metrics," in *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2017, pp. 4277–4281.