Face Photo-Sketch Synthesis via Knowledge Transfer

Mingrui Zhu[1,2] , **Nannan Wang**[1,3*] , **Xinbo Gao**[1,2] , **Jie Li**[1,2] and **Zhifeng Li**[4]
[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China
[2]School of Electronic Engineering, Xidian University, Xi'an, China
[3]School of Telecommunications Engineering, Xidian University, Xi'an, China
[4]Tencent AI Lab, Shenzhen, China
mrz.edu@gmail.com, nnwang@xidian.edu.cn, {xbgao,leejie}@mail.xidian.edu.cn,
michaelzfli@tencent.com

Abstract

Despite deep neural networks have demonstrated strong power in face photo-sketch synthesis task, their performance, however, are still limited by the lack of training data (photo-sketch pairs). Knowledge Transfer (KT), which aims at training a smaller and fast student network with the information learned from a larger and accurate teacher network, has attracted much attention recently due to its superior performance in the acceleration and compression of deep neural networks. This work has brought us great inspiration that we can train a relatively small student network on limited training data by transferring knowledge from a larger teacher model trained on enough training data for other tasks. Therefore, we propose a novel knowledge transfer framework to synthesize face photos from face sketches or synthesize face sketches from face photos. Particularly, we utilize two teacher networks trained on large amount of data in related task to learn knowledge of face photos and knowledge of face sketches separately and transfer them to two student networks simultaneously. The two student networks, one for $photo \rightarrow sketch$ task and the other for $sketch \rightarrow photo$ task, can mimic and transform two kind of knowledge and transfer their knowledge mutually. With the proposed method, we can train a model which has superior performance using a small set of photo-sketch pairs. We validate the effectiveness of our method across several datasets. Quantitative and qualitative evaluations illustrate that our model outperforms other state-of-the-art methods in generating face sketches (or photos) with high visual quality and recognition ability.

1 Introduction

Sketch is one of the most fundamental artistic drawing style which can intuitively and gracefully record scene with monochrome lines. Lots of people want to draw their face sketches for collection or sharing on social media platforms but are limited by insufficient time or their poor artistic skills. In law enforcement and criminal cases, face photos of suspects are sometimes not well recorded by surveillance cameras. In this condition, face sketches of suspects based on the descriptions of eyewitnesses are often drawn by experienced artists as alternatives. However, due to the great gap in texture appearance between face photos and sketches, it is hard to accurately retrieve the suspect in the police mug shot database based on his face sketch. Therefore, face photo-sketch synthesis, which aims at synthesizing a face sketch (photo) from a face photo (sketch) automatically, has attracted much attention and achieved wide applications in both digital entertainment and law enforcement.

Much effort has been devoted to this topic. Exemplar based methods [Wang *et al.*, 2014], started by [Tang and Wang, 2003], had been mainstream in this field until 2015. These methods mining correspondences between input image (image patch) and images (image patches) in a reference set of photo-sketch pairs. The output image is directly reconstructed by the combination of sample images or image patches in reference set. Two main drawbacks often limit their performance: 1) blurry or over smooth, in other words, not photo/sketch-realistic; 2) time-consuming. In recent years, Deep Neural Networks (DNN), especially Generative Adversary Networks (GAN) [Goodfellow *et al.*, 2014], have renewed the state-of-the-art performance in this field. The end-to-end property gives them the advantage of high computational efficiency. However, relatively shallow Convolutional Neural Networks (CNN) are unable to model the highly nonlinear mapping between face photos and face sketches well, resulting in poor visual quality of their generated images. GAN based methods [Wang *et al.*, 2018; Yu *et al.*, 2018] have capacity to generate images with realistic textures which benefit from the adversarial loss. However, undesirable artifacts and mode collapse are often their pain points. Besides, it is known that the desirable performance of deep neural networks depending on large-scale training data. Therefore, lacking of training data (photo-sketch pairs) extremely limit the performance of DNN based methods.

In this paper, we propose a knowledge transfer framework for face photo-sketch synthesis that performs well on a small amount of training data and can generate photo/sketch-realistic results. Our inspiration comes from a thriving research field knowledge transfer [Hinton *et al.*, 2014] recently.

[*]Corresponding Author: Nannan Wang

The basic idea of knowledge transfer is to accelerate and compress the deep neural networks by transferring knowledge from a large teacher model into a small one by learning the class distributions provided by the teacher via softened softmax. Our task differs from the original knowledge transfer in two aspects: 1) our objective is reconstruction rather than classification; 2) we have no enough training data to train a large teacher model. Therefore, we design a special network architecture to novelly combine with knowledge transfer for reconstruction task. Specifically, we utilize two VGG-19 [Simonyan and Zisserman, 2014] trained on ImageNet [Deng *et al.*, 2009] for object recognition task as teacher models to extract knowledge of face photos and knowledge of face sketches separately and transfer them to two student networks simultaneously. The two student networks, one for $photo \rightarrow sketch$ task and the other for $sketch \rightarrow photo$ task, can mimic and transform two kind of knowledge and transfer their knowledge mutually.

To summarize, the contributions of this work are as follows:

- To the best of our knowledge, our method is the first work that introduce knowledge transfer for face photo-sketch synthesis task to tackle the problem of insufficient training data.
- We design a novel network architecture which allow us to transfer knowledge from two teacher models to two student models and transfer knowledge between two student models mutually.
- We conduct extensive experiments to verify the effectiveness of our method. Both qualitative and quantitative results demonstrate the superiority of our method compared with other state-of-the-art methods.

2 Related Work

2.1 Face Photo-Sketch Synthesis

Exemplar based methods can be categorized into three classes: 1) subspace learning-based approaches [Tang and Wang, 2003; Liu *et al.*, 2005]; 2) sparse representation-based approaches [Liang Chang *et al.*, 2010]; and 3) Bayesian inference-based approaches [Tang and Wang, 2009; Zhou *et al.*, 2012; Zhu *et al.*, 2017b]. Due to paper space limitation, we would not review exemplar based methods in this work. A detailed overview of existing exemplar based methods can be found in [Zhu *et al.*, 2017b]. Blurring effect and time consuming are two drawbacks that perplex exemplar based methods.

Deep neural networks has brought great vitality to this field in recent years. Zhang *et al.* [2015] proposed an end-to-end Fully Convolutional Networks (FCN) to directly model the nonlinear mapping between face photos and sketches. However, limited by shallow layers and intensity based Mean Square Error (MSE) loss, it failed to capture texture details and produced undesirable results. Generative adversarial networks [Goodfellow *et al.*, 2014] has provided a power tool to image generation task. It is known to be adept in generating realistic images with crisp textures. Isola *et al.* [2017] proposed a general purpose model named ”pix2pix” which uses conditional GAN for several image to image translation tasks such as labels to street scenes, edges to photos, day to night, etc. Zhu *et al.* [2017a] extended this model by introducing a novel cycle consistency loss. Their ”CycleGAN” model can train in the absence of paired training data and translate image between two domains. Wang *et al.* [2018] drew on this model to face photo-sketch synthesis task and proposed to use discriminator on multiple scales. Yu *et al.* [2018] proposed a Composition-Aided Generative Adversarial Network (CA-GAN) for face photo-sketch synthesis by introducing additional facial composition information to conditional GAN. These methods have capacity to generate images with realistic textures. However, undesirable artifacts and mode collapse are often their pain points. Besides, it is known that the desirable performance of deep neural networks depending on large-scale training data. Therefore, lacking of training data (photo-sketch pairs) extremely limit the performance of these methods. To tackle this problem, Chen *et al.* [2018] proposed a semi-supervised learning framework which can augment paired training samples by synthesizing pseudo sketch features of additional training photos with the help of a small reference set of photo-sketch pairs. This strategy successfully enhances the generalization ability of the model for face photos in the wild and enables the model to generate visual pleasing results. However, sketches generated by this model look over smooth and somewhat different from ground truth sketches which result in the loss of identity information.

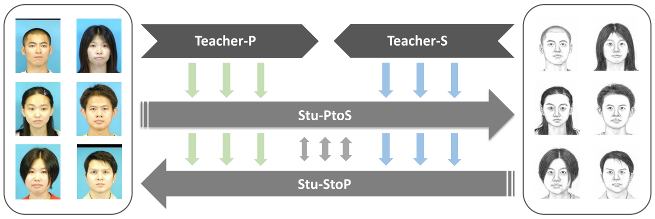

Figure 1: Concept of the proposed knowledge transfer framework.

2.2 Knowledge Transfer

Knowledge Distillation (KD) [Hinton *et al.*, 2014] is the pioneering work to apply knowledge transfer to deep neural networks. A softened version of the final output of a teacher network is used to transfer information to a small student network. With this strategy, a small network can learn how a large network studied given tasks in a compressed form. Attention Transfer [Zagoruyko and Komodakis, 2017] and Neuron Selectivity Transfer [Huang and Wang, 2017] explored different forms of knowledge based on intermediate feature maps of CNNs to improve the performance. Xu *et al.* [2018] proposed a GAN-based approach to learn the loss function to transfer knowledge from teacher to student. A more detailed overview of KD can be found in [Xu *et al.*, 2018]. We borrow ideas from this field and design a novel knowledge transfer framework for face photo-sketch synthesis task.

3 Method

In this section, details of the proposed knowledge transfer framework are presented. We first explain our inspirations

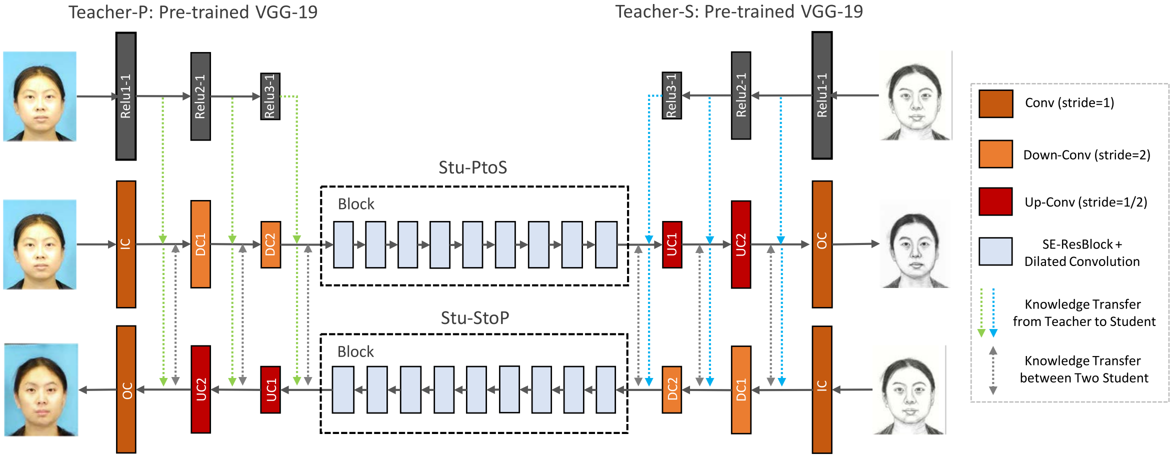

Figure 2: Network architecture of the proposed knowledge transfer framework.

for the knowledge transfer framework. Then, detailed description of the network architecture is presented. Finally, objective functions of the proposed model are provided.

3.1 Knowledge Transfer Framework

Given paired training face photo-sketch samples $\{(x_i, y_i)|x_i \in X, y_i \in Y\}_{i=1}^{N}$, our task is to learn a mapping function from X to Y and a mapping function from Y to X. Because paired training face photo-sketch samples are lacking in the task, we are not able to train a large and accurate network. How to train desirable networks with insufficient training data? Our insight is to learn knowledge from large and accurate teacher networks trained for related tasks. Figure 1 illustrate the concept of the proposed knowledge transfer framework. We utilize two teacher models (Teacher-P and Teacher-S) to distill knowledge of face photos and knowledge of face sketches separately and transfer them to two student networks (Stu-PtoS and Stu-StoP) simultaneously. Stu-PtoS learns the mapping function from photo to sketch under the guidance of Teacher-P and Teacher-S. Similarly, Stu-StoP learns the mapping function from sketch to photo under the guidance of Teacher-S and Teacher-P. In addition, Stu-PtoS and Stu-StoP can transfer and consolidate their knowledge mutually.

3.2 Network Architecture

The architecture of the proposed knowledge transfer framework is presented in Figure 2. In this work, we utilize VGG-19 [Simonyan and Zisserman, 2014] trained on ImageNet [Deng *et al.*, 2009] for object recognition task as teacher model. Intermediate feature maps distilled from Relu1-1, Relu2-1 and Relu3-1 layers of Teacher-P are used as knowledge of face photo domain. Intermediate feature maps distilled from Relu3-1, Relu2-1 and Relu1-1 layers of Teacher-S are used as knowledge of face sketch domain. Stu-PtoS and Stu-StoP have the same structure but the opposite direction. For each student network, we use the downsampling and upsampling structures with nine modified residual blocks between them. The original residual block proposed in [He *et al.*, 2016] has proved good performance for image-to-image translation task in [Zhu *et al.*, 2017a]. Hu *et al.* [2018] proposed a Squeeze-and-Excitation (SE) block and further combined it with residual block to form SE-ResBlock. This block can adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels. We use dilated convolution with dilatation rate 2 in SE-ResBlock to expand its reception field. Eventually, the modified residual block has better performance than the original residual block. The configuration of each student network is specified as follows: IC-k7s1n64, DC1-k3s2n128, DC2-k3s2n256, Block, $\mathrm{UC1-k3s\frac{1}{2}n128}$, $\mathrm{UC2-k3s\frac{1}{2}n64}$, $\mathrm{OC-k7s1n3}$, where k denotes kernel size, s denotes stride, n denotes kernel numbers and Block denotes nine modified residual blocks. Therefore, knowledge distilled from Teacher-P: Relu1-1, Teacher-P: Relu2-1, Teacher-P: Relu3-1, Teacher-S: Relu3-1, Teacher-S: Relu2-1, Teacher-S: Relu1-1 is transferred to Stu-PtoS: IC, Stu-PtoS: DC1, Stu-PtoS: DC2, Stu-PtoS: Block, Stu-PtoS: UC1, Stu-PtoS: UC2 in sequence and transferred to Stu-StoP: UC2, Stu-StoP: UC1, Stu-StoP: Block, Stu-StoP: DC2, Stu-StoP: DC1, Stu-StoP: IC in sequence. Moreover, knowledge distilled from corresponding layers of two student models is transferred mutually.

3.3 Objective

For each paired training face photo-sketch sample $\{(x, y)|x \in X, y \in Y\}$, we denote intermediate feature maps distilled from Teacher-P and Teacher-S as $\phi^j(x, y), j \in \Xi(\mathrm{Teacher-P:Relu1-1, Teacher-P:Relu2-1, Teacher-P:Relu3-1, Teacher-S:Relu3-1, Teacher-S:Relu2-1, Teacher-S:Relu1-1})$, denote intermediate feature maps distilled from Stu-PtoS as $G^m(x), m \in \Upsilon(\mathrm{Stu-PtoS:IC, Stu-PtoS:DC1, Stu-PtoS:DC2, Stu-PtoS:Block, Stu-PtoS:UC1, Stu-PtoS:UC2})$, denote intermediate feature maps distilled from Stu-StoP as $F^n(y), n \in \Gamma(\mathrm{Stu-StoP:UC2, Stu-StoP:UC1, Stu-StoP:Block, Stu-StoP:DC2, Stu-StoP:DC1, Stu-StoP:IC})$, denote the final output of Stu-PtoS as $G(x)$, denote the final output of Stu-StoP as $F(y)$. Our objective contains three terms: teacher-student knowledge transfer loss; student-student knowledge transfer loss; and image reconstruction Loss.

By introducing L2 distance to prevent the intermediate feature maps distilled from teacher model and student model contradicting each other, the teacher-student knowledge transfer loss can be expressed as follows:

$$\begin{aligned}\mathcal{L}_{tea-stu}(x,y) = &\sum_{j\in\Xi,m\in\Upsilon} \parallel \phi^j(x,y) - G^m(x) \parallel_2^2 /C_jH_jW_j \\ &+ \sum_{j\in\Xi,n\in\Gamma} \parallel \phi^j(x,y) - F^n(y) \parallel_2^2 /C_jH_jW_j\end{aligned} \tag{1}$$

where C_j, H_j and W_j indicate channel numbers, height and width of intermediate feature maps of layer j, respectively.

By introducing L2 distance to prevent the intermediate feature maps distilled from two student models contradicting each other, the student-student knowledge transfer loss can be expressed as follows:

$$\begin{aligned}&\mathcal{L}_{stu-stu}(x,y) = \\ &\sum_{m\in\Upsilon,n\in\Gamma} \parallel G^m(x) - F^n(y) \parallel_2^2 /C_mH_mW_m\end{aligned} \tag{2}$$

where C_m, H_m and W_m indicate channel numbers, height and width of intermediate feature maps of layer m, respectively.

By introducing L2 distance to prevent the final output of each student model and its ground truth contradicting each other, the image reconstruction loss can be expressed as follows:

$$\mathcal{L}_{rec}(x,y) = \parallel y - G(x) \parallel_2^2 + \parallel x - F(y) \parallel_2^2 \tag{3}$$

By combining above losses, we can achieve our full objective:

$$\mathcal{L}_{total} = \mathcal{L}_{tea-stu} + \mathcal{L}_{stu-stu} + \mathcal{L}_{rec} \tag{4}$$

In the training stage, we alternately train Stu-PtoS and Stu-StoP for each iteration.

4 Experiments

In this section, we first introduce the experimental settings. To further validate the effectiveness of our method, We conduct ablation studies to verify the effectiveness of various components and compare our results with state-of-the-art methods. Both visual and quantitative comparisons demonstrate the superiority of our method.

4.1 Experimental Settings

We conducted experiments on two public datasets: the CUFS dataset [Tang and Wang, 2009] and the CUFSF dataset [Zhang *et al.*, 2011b]. The CUFS dataset consists of 188 faces from the Chinese University of Hong Kong (CUHK) student database [Tang and Wang, 2003], 123 faces from the AR database [Martinez and Benavente, 1998], and 295 faces from the XM2VTS database [Messer *et al.*, 1999]. There is a photo-sketch pair for each face under normal lighting condition, and with a neutral expression. The CUFSF dataset consists of 1,194 faces from the FERET database [Phillips *et al.*, 2000]. For each face, there is a photo with lighting variation and a sketch with shape exaggeration drawn by an artist. Therefore, this dataset is more challenging than the CUFS dataset. All images are processed by aligning center of the eyes to the fixed position and cropping to the size of 200 × 250. For the CUHK student database, we randomly choose 88 face photo-sketch pairs for training and the rest are used for testing. For the AR database, 80 face photo-sketch pairs are randomly chosen for training and the rest 43 pairs are used for testing. For the XM2VTS database, we randomly choose 100 pairs for training and the rest 195 pairs are used for testing. For the CUFSF dataset, 250 face photo-sketch pairs are chosen for training and the rest are used for testing.

Our model was trained on a NVIDIA Titan X GPU. Adam [Kingma and Ba, 2015] with $\beta_1 = 0.5$ was used for optimization. The learning rate was set to 0.0002 and the iteration times was set to 200. Weights were initialized from a Gaussian distribution with mean 0 and standard deviation 0.02. We scaled the size of the input images to 256 × 256 and normalized the pixel value to the interval $[-1, 1]$ before putting them into the model. The number of input and output channels was set to 3. We updated Stu-PtoS and Stu-StoP alternatively at every iteration. The batch size was set to 1.

We utilized Feature SIMilarity index (FSIM) [Zhang *et al.*, 2011a] to quantitatively evaluate our model. Structural Similarity Index Metric (SSIM) [Wang *et al.*, 2004] was frequently used to evaluate the performance of exemplar based methods. However, we find that it is not suitable for evaluating deep learning models. One phenomena is that synthesized images which look more blurry tends to have higher SSIM score. This is contrary to human visual perception which biases to sharper images. FSIM captures similarity between low-level features of the synthesized image and the ground-truth image, which has much higher consistency with human visual perception. Therefore, we use FSIM instead of SSIM as criterion to evaluate the performance of deep learning models.

Null-space Linear Discriminant Analysis (NLDA) [Chen *et al.*, 2000] was utilized to evaluate the face recognition accuracy of the synthesized sketches on the CUFSF dataset. We randomly choose 300 synthesized sketches and their corresponding sketches drawn by the artist for classifier training. The rest 644 synthesized sketches are used as the gallery set and their corresponding sketches drawn by the artist are used as query images. We repeated each face recognition experiment 50 times by randomly partition the data.

4.2 Ablation Study

To demonstrate the advantage of knowledge transfer strategy, we compare the results based on the following configurations on the CUHK student database: (a) using only image reconstruction loss; (b) using image reconstruction loss and teacher-student knowledge transfer loss; and (c) using image reconstruction loss, teacher-student knowledge transfer loss and student-student knowledge transfer loss. Figure 3 shows some synthesized photos and sketches based on the

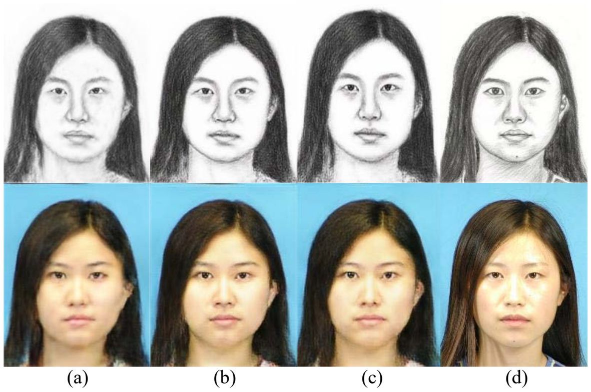

Figure 3: Results of ablation study: (a) only image reconstruction loss; (b) image reconstruction loss + teacher-student knowledge transfer loss; (c) image reconstruction loss + teacher-student knowledge transfer loss + student-student knowledge transfer loss; and (d) Ground truth.

Configurations	Photo	Sketch
$\mathcal{L}_{rec}$	0.7778	0.7403
$\mathcal{L}_{rec} + \mathcal{L}_{tea-stu}$	0.7803	0.7485
$\mathcal{L}_{rec} + \mathcal{L}_{tea-stu} + \mathcal{L}_{stu-stu}$	**0.7835**	**0.7491**

Table 1: Average FSIM socre of the synthesized face photos/sketches on the CUHK student database based on different configurations.

above configurations on the CUHK student dataset. It can be observed that the visual quality of the results improves significantly with the help of teacher-student knowledge transfer loss. By adding student-student knowledge transfer loss, the visual quality further improves slightly. Table 1 shows the average FSIM score of the synthesized photos/sketches on the CUHK student database based on different configurations. The evaluation results are consistent with our visual observations. To further validate the effectiveness of the knowledge transfer strategy, we visualize the intermediate feature maps of corresponding layers of all teacher and student models with or without knowledge transfer. As shown in Figure 4, the intermediate feature maps of student models have great difference in appearance with teacher models without knowledge transfer. With the help of knowledge transfer, the intermediate feature maps of corresponding layers of all teacher and student models have consistent appearance. This indicate that the two student models have learnt the knowledge from teachers and from each other.

4.3 Comparison with State-of-the-art Methods

We compare our method with six state-of-the-art methods: DGFL [Zhu *et al.*, 2017b], FCN [Zhang *et al.*, 2015], pix2pix [Isola *et al.*, 2017], CycleGAN [Zhu *et al.*, 2017a], PS2MAN [Wang *et al.*, 2018] and FaceSketchWild [Chen *et al.*, 2018]. Owing to spatial confined, we only compare our method with one exemplar based method DGFL which has shown the best performace and focus on the comparison with deep learning models. All results are obtained from the source codes provided by the authors except the results of FCN. We implement FCN by ourselves and get the results which are consistent with the original work. Because DGFL, FCN and FaceSketchWild methods are designed for face $photo \rightarrow sketch$ synthesis task, we only use their synthesized face sketches. Other methods have both synthesized face photos and face sketches.

Figure 5 presents some synthesized face photos from different methods on the CUFS dataset and the CUFSF dataset.

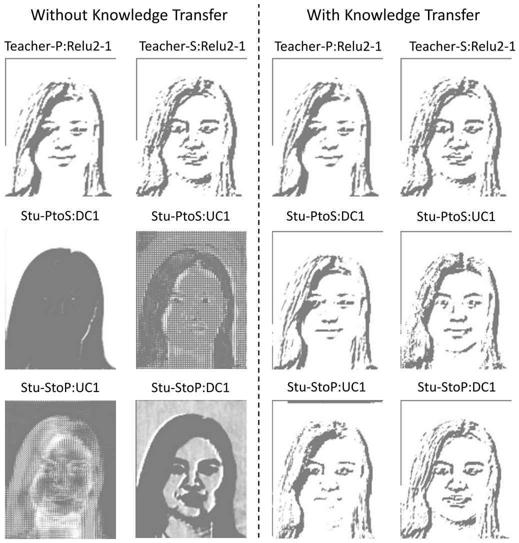

Figure 4: Intermediate feature maps of corresponding layers of all teacher and student models with or without knowledge transfer.

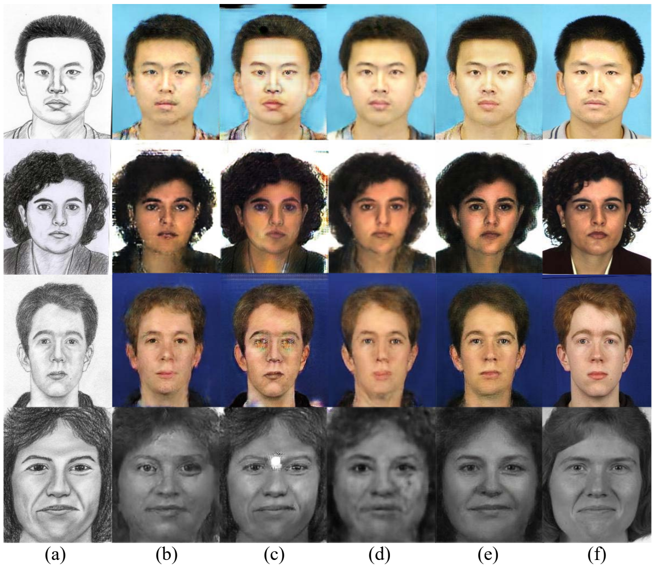

Figure 5: Examples of synthesized face photos on the CUFS dataset and the CUFSF dataset. (a) Input, (b) pix2pix, (c) CycleGAN, (d) PS2MAN, (e) the proposed method, and (f) Ground Truth. From top to bottom, the examples are selected from the CUHK student database, the AR database, the XM2VTS database, and the CUFSF database, sequentially.

		DGFL	FCN	pix2pix	CycleGAN	PS2MAN	FaceSketchWild	KnowledgeTransfer
Photo	CUFS	-	-	0.7726	0.7450	0.7819	-	**0.7851**
	CUFSF	-	-	0.7777	0.7645	0.7812	-	**0.7931**
Sketch	CUFS	0.7079	0.6936	0.7363	0.7219	0.7230	0.7114	**0.7373**
	CUFSF	0.6957	0.6624	0.7283	0.7088	0.7233	0.6821	**0.7311**

Table 2: Average FSIM score of the synthesized face photos/sketches on the CUFS dataset and the CUFSF dataset based on different methods.

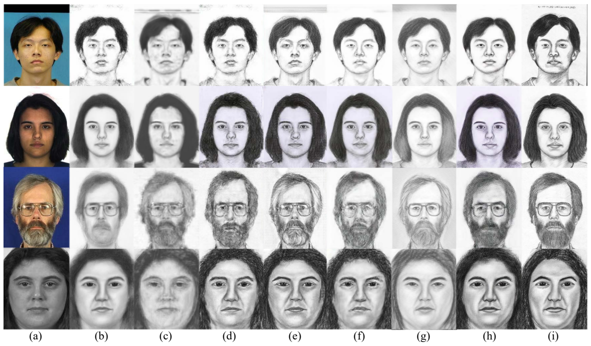

Figure 6: Examples of synthesized face sketches on the CUFS dataset and the CUFSF dataset. (a) Input, (b) DGFL, (c) FCN, (d) pix2pix, (e) CycleGAN, (f) PS2MAN, (g) FaceSketchWild, (h) the proposed method, and (i) Ground Truth. From top to bottom, the examples are selected from the CUHK student database, the AR database, the XM2VTS database, and the CUFSF database, sequentially.

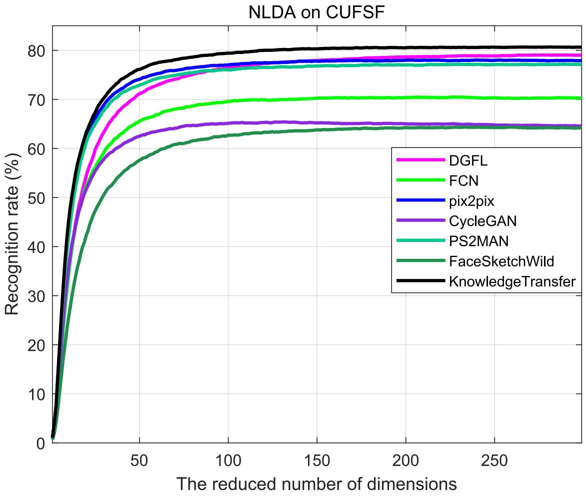

Figure 7: NLDA face recognition accuracies with the variation of the reduced number of dimensions of different methods on the CUFSF dataset.

The results of pix2pix and CycleGAN have sharp edges but possess obvious artifacts and noise. PS2MAN produces less artifacts but its results are blurry. By comparison, our method can synthesize clear face photos with reasonable texture.

Some synthesized face sketches from different methods on the CUFS dataset and the CUFSF dataset are shown in Figure 6. The results of DGFL and FCN are blurry. GAN based methods (pix2pix, CycleGAN and PS2MAN) can generate sketch-like textures. However, some undesirable noises are produced in eye and nose areas. FaceSketchWild has stronger robustness to the environment noises but tends to generate over smooth results. By comparison, our method can generate the most sketch-like textures while generate the least noises.

Table 2 presents the average FSIM score of the synthesized face photos/sketches on the CUFS dataset and the CUFSF dataset based on different methods. Obviously, our method obtains the best FSIM values.

Due to space limitation, we only present the NLDA face recognition accuracies with the variation of the reduced number of dimensions of different methods on the more challenging CUFSF dataset. As shown in Figure 7, the proposed method achieve the best performance.

5 Conclusion

In this paper, we propose a knowledge transfer framework for face photo-sketch synthesis task. By transferring knowledge from a larger and accurate teacher model trained on enough training data for other tasks, we can train a relatively small but desirable student network on limited training data using the proposed framework. In addition, knowledge transfer between two student models can further improve the performance slightly. We compare the proposed model with recent state-of-the-art methods on two public datasets. Both qualitative and quantitative results demonstrate that the proposed method achieves significant improvements. In our future work, we plan to explore more meaningful and suitable knowledge form for the knowledge transfer framework.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (under Grant 61876142, 61432014, U1605252, 61772402, and 61671339), in part by the National Key Research and Development Program of China under Grant 2016QY01W0200, in part by the National High-Level Talents Special Support Program of China under Grant CS31117200001, in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2016QNRC001, in part by the Young Talent fund of University Association for Science and Technology in Shaanxi, China, in part by the CCF-Tencent Open Research Fund (No. RAGR20180105) and Tencent AI Lab Rhino-Bird Focused Research Program (No. JR201923), and in part by the Xidian University-Intellifusion Joint Innovation Laboratory of Artificial Intelligence, in part by the Fundamental Research Funds for the Central Universities under Grant JB190117.

References

[Chen *et al.*, 2000] Li-Fen Chen, Hong-Yuan Mark Liao, Ming-Tat Ko, Ja-Chen Lin, and Gwo-Jong Yu. A new lda-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33(10):1713–1726, 2000.

[Chen *et al.*, 2018] Chaofeng Chen, Wei Liu, Xiao Tan, and Kwan-Yee K. Wong. Semi-supervised learning for face sketch synthesis in the wild. In *ACCV*, 2018.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hinton *et al.*, 2014] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Workshop*, 2014.

[Hu *et al.*, 2018] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. In *CVPR*, 2018.

[Huang and Wang, 2017] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv:1707.01219*, 2017.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[Liang Chang *et al.*, 2010] Mingquan Zhou Liang Chang, Yanjun Han, and Xiaoming Deng. Face sketch synthesis via sparse representation. In *ICPR*, pages 2146–2149, 2010.

[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 *CVPR*, pages 1005–1010, 2005.

[Martinez and Benavente, 1998] A. M. Martinez and Robert Benavente. The ar face database. Technical report, CVC Technical Report #24, 1998.

[Messer *et al.*, 1999] Kieron Messer, Jiri Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *AVBPA*, pages 72–77, 1999.

[Phillips *et al.*, 2000] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The feret evaluation methodology for face recognition algorithms. *TPAMI*, 22(10):1090–1104, 2000.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[Tang and Wang, 2003] Xiaoou Tang and Xiaogang Wang. Face sketch synthesis and recognition. In *ICCV*, pages 687–694, 2003.

[Tang and Wang, 2009] Xiaoou Tang and Xiaogang Wang. Face photo-sketch synthesis and recognition. *TPAMI*, 31(11):1955–1967, November 2009.

[Wang *et al.*, 2004] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *TIP*, 13(4):600–612, 2004.

[Wang *et al.*, 2014] Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. A comprehensive survey to face hallucination. *IJCV*, 106(1):9–30, January 2014.

[Wang *et al.*, 2018] Lidan Wang, Vishwanath A. Sindagi, and Vishal M. Patel. High-quality facial photo-sketch synthesis using multi-adversarial networks. In *FG 2018*, pages 83–90, 2018.

[Xu *et al.*, 2018] Zheng Xu, Yen-Chang Hsu, and Jiawei Huang. Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks. *arXiv:1709.00513v2*, 2018.

[Yu *et al.*, 2018] Jun Yu, Shengjie Shi, Fei Gao, Dacheng Tao, and Qingming Huang. Composition-aided face photo-sketch synthesis. *arXiv:1712.00899v3*, 2018.

[Zagoruyko and Komodakis, 2017] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

[Zhang *et al.*, 2011a] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *TIP*, 20(8):2378–2386, 2011.

[Zhang *et al.*, 2011b] Wei Zhang, Xiaogang Wang, and Xiaoou Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*, pages 513–520, 2011.

[Zhang *et al.*, 2015] Liliang Zhang, Liang Lin, Xian Wu, Shengyong Ding, and Lei Zhang. End-to-end photo-sketch generation via fully convolutional representation learning. In *ACM ICMR*, pages 627–634, 2015.

[Zhou *et al.*, 2012] Hao Zhou, Zhanghui Kuang, and Kwan-Yee K. Wong. Markov weight fields for face sketch synthesis. In *CVPR*, pages 1091–1097, 2012.

[Zhu *et al.*, 2017a] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, pages 2223–2232, 2017.

[Zhu *et al.*, 2017b] Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Li. Deep graphical feature learning for face sketch synthesis. In *IJCAI*, pages 3574–3580, 2017.