A Sketch-Transformer Network for Face Photo-Sketch Synthesis

Mingrui Zhu[1], **Changcheng Liang**[1], **Nannan Wang**[1*], **Xiaoyu Wang**[2], **Zhifeng Li**[3] and **Xinbo Gao**[4]

[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China
[2]The Chinese University of Hong Kong (Shenzhen), Shenzhen, China
[3]Tencent, Shenzhen, China
[4]Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing, China
mrzhu@xidian.edu.cn, ccliang@stu.xidian.edu.cn, nnwang@xidian.edu.cn, fanghuaxue@gmail.com, michaelzfli@tencent.com, gaoxb@cqupt.edu.cn

Abstract

We present a face photo-sketch synthesis model, which converts a face photo into an artistic face sketch or recover a photo-realistic facial image from a sketch portrait. Recent progress has been made by convolutional neural networks (CNNs) and generative adversarial networks (GANs), so that promising results can be obtained through real-time end-to-end architectures. However, convolutional architectures tend to focus on local information and neglect long-range spatial dependency, which limits the ability of existing approaches in keeping global structural information. In this paper, we propose a Sketch-Transformer network for face photo-sketch synthesis, which consists of three closely-related modules, including a multi-scale feature and position encoder for patch-level feature and position embedding, a self-attention module for capturing long-range spatial dependency, and a multi-scale spatially-adaptive de-normalization decoder for image reconstruction. Such a design enables the model to generate reasonable detail texture while maintaining global structural information. Extensive experiments show that the proposed method achieves significant improvements over state-of-the-art approaches on both quantitative and qualitative evaluations.

Figure 1: A comparision of face photo sketch synthesis results between the proposed Sketch-Transformer and a state-of-the-art (SOTA) approach. Sketch-Transformer (ours) can capture long-range spatial dependency while generate reasonable detail texture.

1 Introduction

Generating a face sketch (photo) from a face photo (sketch), often referred as face photo-sketch synthesis, is an important task in computer vision. It has many applications in digital entertainment, animation production and law enforcement [Wang *et al.*, 2014; Li *et al.*, 2016]. The core challenge of face photo-sketch synthesis lies in synthesizing visually realistic and semantically plausible images and surpassing the considerable discrepancies (shape, texture and color) barrier.

Early studies [Liu *et al.*, 2005; Liang Chang *et al.*, 2010; Zhu *et al.*, 2017b] attempt to solve the problem in an exemplar-based manner, i.e. matching and combining sample images (image patches) in a reference set of photo-sketch pairs to synthesize the target image. These approaches work well under constrained conditions such as less illumination variations, pose changes, and deformations, but will fail when come across more complicate conditions. Moreover, two main flaws often limit their performance: 1) blurry or over smooth, i.e not realistic; 2) time-consuming. Rapid progress in deep convolutional neural networks (CNN), especially in

*Corresponding Author: Nannan Wang

generative adversarial networks (GAN) [Goodfellow *et al.*, 2014], has inspired recent studies [Wang *et al.*, 2018; Yu *et al.*, 2020; Chen *et al.*, 2018] to formulate face photo sketch synthesis as a image-to-image translation [Isola *et al.*, 2017; Zhu *et al.*, 2017a] problem. With the assistance of the adversarial loss, these approaches have capacity to generate images with realistic textures.

Although promising results have been obtained, the intrinsic shortage of convolutional architectures that lacks of the ability of capturing long-range spatial dependency has limited the performance of existing approaches, which may results in the loss of global structure information and thus generating images with compromised visual quality. As shown in Figure 1, the results of a state-of-the-art (SOTA) method [Yu *et al.*, 2020] have undesirable artifacts and distorted structures. Recently, transformer models [Vaswani *et al.*, 2017] which mainly based on self-attention mechanism have demonstrated exemplary performance on natural language processing (NLP) tasks and intrigued the vision community to investigate their application to computer vision problems [Dosovitskiy *et al.*, 2020]. Inspired by the power of transformer in NLP and many computer vision tasks, we investigate its application in face photo-sketch synthesis task in this work. However, there are three factors that limit the application of existing transformer models in this task: 1) The training samples are limited so that the model should not be too large; 2) The resolution of the image is relatively large so that the self-attention module consumes lots of computing resources; 3) The self-attention module is unable to capture positional information of the tokens in an image.

To address these problems, we propose a Sketch-Transformer which can properly introduce the self-attention mechanism into the face photo-sketch synthesis task. Specifically, three closely-related modules are proposed. First, we propose a multi-scale feature and position encoder (MFP-Encoder) which integrates convolutional architectures and a face parsing model to extract multi-scale feature embeddings and positional encodings in each local area. Second, we stack several residual self-attention layers in the bottleneck to capture the long-range spatial dependency between the tokens (local embeddings). Finally, we propose a multi-scale spatially-adaptive de-normalization decoder (MSPADE-Decoder) which takes as input the output of the self-attention module, multi-scale feature embeddings and positional encodings generated by the multi-scale feature and position encoder to reconstruct the target image. The overall design enables our Sketch-Transformer to capture long-range spatial dependency while generate reasonable detail texture and therefore achieve a better visual result compared with state-of-the-art approaches (as shown in Figure 1).

The contributions of this work are summarized as follows:

- We propose to learn the key elements of the transformer architecture and adapt them to face photo-sketch synthesis task.
- We propose a Sketch-Transformer with three closely-related modules to properly introduce the self-attention mechanism. The proposed model can capture long-range spatial dependency while generate reasonable detail texture.
- Quantitative and qualitative experiments demonstrate that the proposed model achieves superior performance compared with other state-of-the-art methods on public benchmarks and face images in real scenarios.

2 Related Work

In this section, we review previous studies of face photo-sketch synthesis and transformer which are the most relevant to our work.

2.1 Face Photo-Sketch Synthesis

Existing works for face photo-sketch synthesis can be mainly divided into two categories. Exemplar-based methods reconstruct target image by mining correspondences between input image (image patch) and images (image patches) in a reference set of photo-sketch pairs. Deep learning-based methods attempt to predict the target image pixels from the source image pixels through an end-to-end convolutional neural networks.

Exemplar-based methods can be further grouped into three types: subspace learning-based approaches [Liu *et al.*, 2005], sparse representation-based approaches [Liang Chang *et al.*, 2010], and Bayesian inference-based approaches [Zhu *et al.*, 2017b]. A detailed overview of existing exemplar-based methods can be found in [Wang *et al.*, 2014].

Recently, CNN-based and GAN-based approaches have emerged as a promising paradigm for face photo-sketch synthesis. Initial effort [Zhang *et al.*, 2015] trains an end-to-end fully convolutional neural networks (FCN) for directly modeling the nonlinear mapping between face photos and face sketches. Limited by shallow layers and pixel-level loss, however, it fails to capture texture details and fails to preserve reasonable structures. Isola et al. [2017] use conditional GAN (cGAN) as a unified solution (pix2pix) for several image-to-image translation tasks such as edges to photos, labels to street scenes, day to night, etc. Zhu et al. [2017a] propose a CycleGAN model for unpaired image-to-image translation by introducing a cycle consistency loss. These two models can be directly applied to face photo-sketch synthesis task. Several works follow ideas from image-to-image translation and focus on improving face photo-sketch synthesis performance by adding prior information. Wang et al. [2018] propose a multi-scale discriminator to provide adversarial supervision on different image resolution. SCAGAN [Yu *et al.*, 2020] introduces facial composition information as additional input to help the generation of sketch portraits and proposes a compositional loss based on facial composition information. To tackle the problem of insufficient paired training data, Chen et al. [2018] propose a semi-supervised learning method to augment paired training samples by synthesizing pseudo sketch features of additional training photos and learn the mapping function between them. Although great progress has been made by above approaches, undesirable artifacts and distorted structures, however, are still exists, especially in the results of real scenarios.

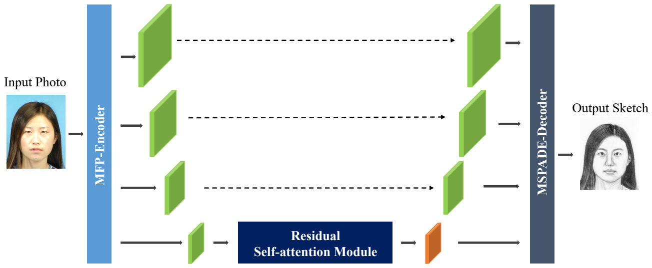

Figure 2: The illustration of the Sketch-Transformer architecture.

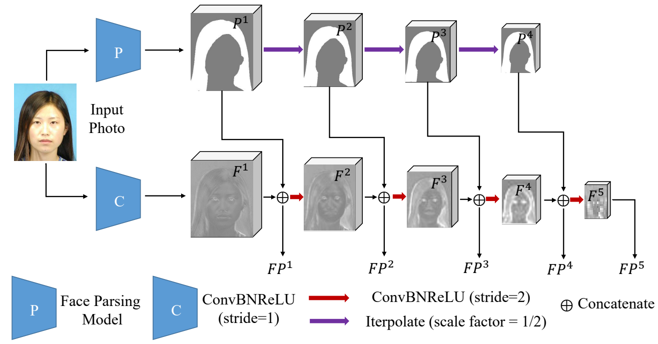

Figure 3: The illustration of the MFP-Encoder.

2.2 Transformer and Self-attention

Transformer is firstly applied on natural language processing (NLP) tasks, which mainly leverages self-attention mechanism to capture long-range dependencies in the input domain. The seminal work of Vaswani et al. [2017] proposes to use solely attention mechanisms for machine translation. Since then, transformer architecture has opened up a new route. Lots of popular methods have been proposed and have achieved the state-of-the-art performance in different NLP tasks. The breakthroughs achieved by transformer in NLP domain have attracted lots of interest in the computer vision community. Many studies have successfully adapted transformer models to varies computer vision tasks including image recognition, object detection, image super-resolution and several other tasks. A comprehensive overview of the vision transformer literature has been introduced by Han et al. [2020].

3 Method

Given paired training face photo-sketch samples $\{(x_i, y_i) \in (X, Y)\}_{i=1}^{N}$, our goal is to learn a mapping function G that maps images from photo domain X to sketch domain Y or learn a mapping function F that maps images from sketch domain Y to photo domain X. The pipeline of the proposed Sketch-Transformer is shown in Figure 2. It consists of three closely-related modules, including a multi-scale feature and position encoder (MFP-Encoder) for patch-level feature and position embedding, a residual self-attention module for capturing long-range spatial dependency, and a multi-scale spatially-adaptive de-normalization decoder (MSPADE-Decoder) for image reconstruction.

3.1 MFP-Encoder

The MFP-Encoder integrates convolutional architectures and a face parsing model to extract multi-scale feature embeddings and positional encodings in each local area. It consists of two paths: a feature embedding path and a position embedding path, as shown in Figure 3.

The feature embedding path utilizes a series of convolution layers (a stride-1 convolution layer and four stride-2 convolution layers) to gradually extract multi-scale features. Therefore, the feature vector of each position in the last activation (FP^5) represents the high-level features of a 16×16 patch in the corresponding local area of the input image. The position embedding path utilizes a face parsing model [Yu *et al.*, 2018] to extract semantic facial labels and scale them to different spatial resolution. Denote semantic facial labels of each layer as M^l, $M^l \in \Re^{c_l \times h_l \times w_l}$, where c_l, h_l, w_l denote component number, height and width of the semantic labels of the l^th feature layer. Each value (0 or 1) in M^l denotes whether the position belongs to the c-th component. Such semantic facial labels actually contain sufficient positional information and can indicate which semantic component the feature embeddings of each position belongs to. We concatenate the feature embeddings and position embeddings at different level to obtain the multi-scale feature and position embeddings. Then, the first four feature and position embeddings are passed to the MSPADE-Decoder as spatial information to help supplement spatial and texture information and the last one is passed to a residual self-attention model to learn long-range dependencies between the embeddings (tokens) from all positions.

3.2 Residual Self-attention Module

Self-attention is the core component of the transformer architecture, which can capture long-range dependency between tokens. From the MFP-Encoder, we obtain the patch-level feature and position embeddings of all positions. However, the relationships between these embeddings are neglect. Therefore, we introduce a residual self-attention module to capture their dependencies. The module consists of nine basic residual self-attention layers. The illustration of each layer is shown in Figure 4.

The intuition behind this module is to update each vector at each position of the embeddings by aggregating global information from all other positions. Through this module, we can get the revised embeddings $\hat{FP}^5$ which have learned the long-range dependencies.

3.3 MSPADE-Decoder

We utilize the spatially-adaptive de-normalization (SPADE) block [Park *et al.*, 2019] on multi-scale feature and position embeddings to gradually reconstruct the target image. More specifically, we utilize positional normalization (PN) [Li *et al.*, 2019] instead of batch normalization (BN) to better preserving the structure information synthesized in prior layers. The illustration of the MSPADE-Decoder is shown in Figure 4.

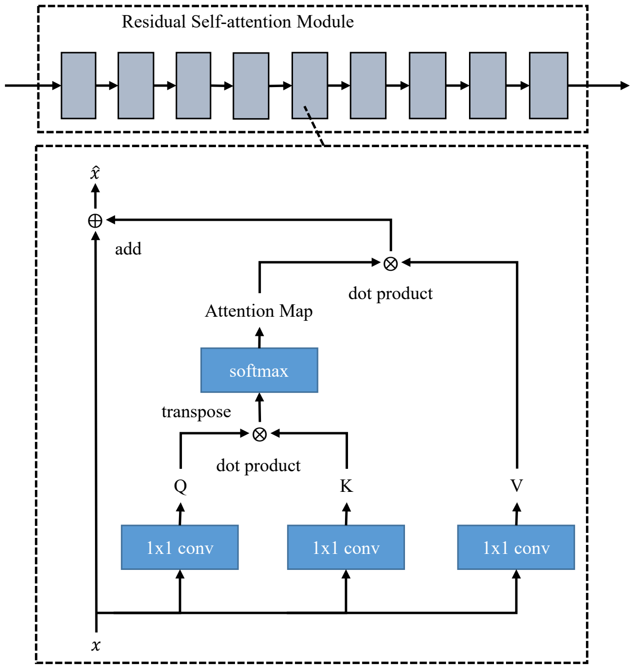

Figure 4: The illustration of the residual self-attention mudule.

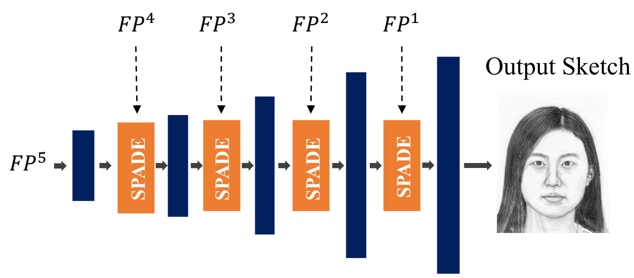

Figure 5: The illustration of the SPADE-Decoder.

3.4 Loss Function

The full loss of our model consists of two loss functions: adversarial loss and perceptual loss. For the sake of brevity, we only describe the losses for photo to sketch synthesis task. The losses for sketch to photo synthesis has the same form. For convenience of expression, we denote the SketchTransformer as G. In order to provide the adversarial loss, we utilize a 70 × 70 PatchGAN discriminator [Isola *et al.*, 2017], which is denoted as D.

Adversarial Loss

In this work, instead of using the vanilla GAN[Goodfellow *et al.*, 2014], we use the Least Squares GAN [Mao *et al.*, 2017] for stable training. For the mapping function G and its discriminator D, we express the objective as:

$$\mathcal{L}_{adversarial} = \mathbb{E}_y[(D_Y(y))^2] + \mathbb{E}_x[(1 - D_Y(G(x)))^2] \quad (1)$$

Perceptual Loss

To ensure that the generated image and its ground truth are similar in semantic feature level, we introduce the perceptual

Database		Training Pairs	Testing Pairs
CUFS	CUHK Student	88	100
	AR	80	43
	XM2VTS	100	195
CUFSF		250	944

Table 1: Partition settings of the databases

loss [Johnson *et al.*, 2016]:

$$\mathcal{L}_{perceptual} = \mathbb{E}_x[\frac{1}{C_j H_j W_j} \parallel \phi_j(G(x)) - \phi_j(y) \parallel_1] \quad (2)$$

where ϕ_j indicates feature maps of the jth layer of a pre-trained VGG-19 model [Simonyan and Zisserman, 2014], C_j, H_j and W_j indicate channel numbers, height and width of the feature maps, respectively.

Full Loss

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

$$\mathcal{L}_{full} = \lambda_1 \mathcal{L}_{adversarial} + \lambda_2 \mathcal{L}_{perceptual} \quad (3)$$

In this work, we set $\lambda_1 = 1$, $\lambda_2 = 5$ to keep corresponding losses in the same order of magnitude.

4 Experiments

In this section, we first discuss the experimental settings. We will then conduct ablation study to quantify the contribution of different configurations to overall effectiveness. Finally, we will compare our results with state-of-the-art methods both qualitatively and quantitatively.

4.1 Implement Details

All models are trained on a NVIDIA Tesla V100 GPU using Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.99$. We train all models with a fixed learning rate of 0.0002 until 300,000 iterations. The batch size is set to 1 for all experiments. 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. During training, we updated G and D alternatively at every iteration.

4.2 Database

The experiments are conducted on two public databases: the CUFS database [Tang and Wang, 2009] and the CUFSF dataset [Zhang *et al.*, 2011b]. The CUFS database consists of 188 identities from the Chinese University of Hong Kong (CUHK) student database [Tang and Wang, 2003], 123 identities from the AR database [Martinez and Benavente, 1998], and 295 identities from the XM2VTS database [Messer *et al.*, 1999]. Each identity has a photo-sketch pair under normal light condition, and with a neutral expression. The CUFSF database has 1,194 identities from the FERET database [Phillips *et al.*, 2000]. For each identity, there is a photo with illumination variation and a sketch with exaggerated structure. Therefore, face photo-sketch synthesis on the CUFSF database is more challenging than on the CUFS dataset. All images are processed by aligning the center of two eyes to the fixed position and cropping to the size of 200×250. The way we divide the training set and the test set is the same as [Zhu *et al.*, 2017b]. For the CUFS database, 88 face photo-sketch pairs in CUHK database, 80 face photo-sketch pairs in AR database and 100 face photo-sketch pairs in XM2VTS database are selected for training and the rest are used for testing. For the CUFSF database, 250 face photo-sketch pairs are selected for training and the rest are used for testing. Table 1 shows the partition settings of the databases.

4.3 Baselines

For fair comparison, we run face photo-sketch synthesis on our method and all baselines for input images of size 200×250 under the same partition setting. We compare our method with seven 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], Wild [Chen *et al.*, 2018] and SCAGAN [Yu *et al.*, 2020]. Among these baselines, DGFL is traditional exemplar-based method which achieves the best performance while the others are deep learning-based methods. 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 FCN, DGFL and Wild methods are designed for face photo $\rightarrow$ sketch synthesis task, we only compare with their synthetic face sketches. Other methods have both synthetic face photos and face sketches that used for comparison.

4.4 Evaluation Metrics

In this paper, we use three types of evaluation metrics to evaluate the objective quality of the synthetic images: the learned perceptual image patch similarity (LPIPS) [Zhang *et al.*, 2018], the Fréchet Inception Distance (FID) [Heusel *et al.*, 2017] and the feature similarity index (FSIM) [Zhang *et al.*, 2011a]. The LPIPS takes two images (image patches) as the input, calculates the L2 distance between their normalized deep feature embeddings, and predicts the perceptual judgment score through the linear layer. A lower score indicates better quality of synthetic images. FID is designed to capture the Fréchet difference of two Gaussians (synthetic and real-world images). We compute the FID score between the synthetic images and real ones. Lower FID score indicates better quality of synthetic images. FSIM is a commonly used metric for full-reference image quality assessment, which captures the similarity between low-level features of images. It shows higher consistency with human visual perception. We calculated the average FSIM score between synthetic images and real ones. A higher FSIM score indicates better quality of synthetic images.

4.5 Ablation Study

We compute the LPIPS (alex) score between the synthetic images and real ones on CUHK database under different configurations to quantify the contribution of different configurations to overall effectiveness. The ablation study is conducted on four configurations: (a) U-net [Ronneberger *et al.*, 2015] architecture; (b) Using MSPADE-Decoder to replace the origin decoder in U-net; (c) Adding residual self-attention module on the basis of (b); (d) Adding position embeddings on the basis of (c) (i.e. Full Sketch-Transformer). The evaluation scores are shown in Table 2, from which we can conclude that all the modifications are critical to the final effectiveness of the proposed method.

Configurations	Photo-LPIPS(alex) $\downarrow$	Sketch-LPIPS(alex) $\downarrow$
(a)	0.1686	0.1732
(b)	0.1529	0.1700
(c)	0.1537	0.1657
(d)	**0.1511**	**0.1662**

Table 2: Ablation study: LPIPS (alex) scores for different variants of configurations, evaluated on CUHK $photo \rightarrow sketch$ and $sketch \rightarrow photo$.

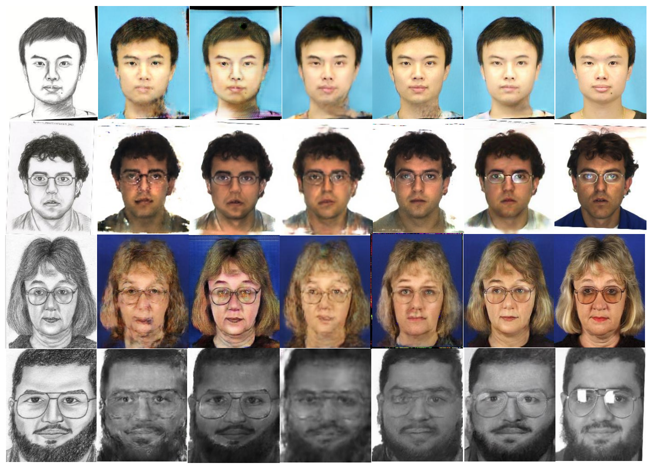

Figure 6: Examples of synthetic face photos on the CUFS dataset and the CUFSF dataset. From top to bottom, the examples are selected from the CUHK student database, the AR database, the XM2VTS database and the CUFSF database, sequentially.

4.6 Comparison with Baselines

Figure 6 presents some synthetic face photos from different methods on the CUFS dataset and 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. Face photos synthesized by the SCAGAN have reasonable texture and less artifacts, but still possess some structure distortions. As shown in the figure, synthetic photos of the proposed method retain considerable structural information and achieve the most reasonable texture distribution, and therefore has the best visual performance.

Some synthetic face sketches from different methods on the CUFS dataset and the CUFSF dataset are shown in Figure 7. The results of DGFL and FCN are too blurry. GAN-based methods (pix2pix, CycleGAN, PS2MAN and SCAGAN) can generate sketch-like textures. However, some undesirable textures are produced in eye and hair areas. Wild

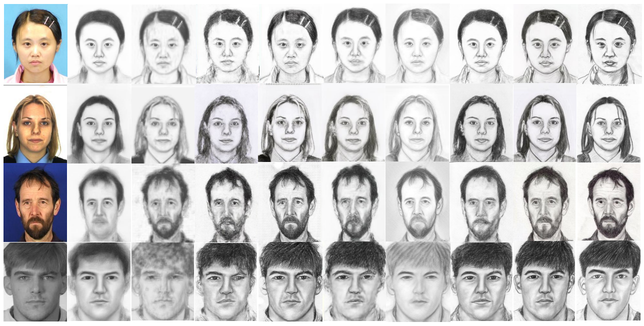

Figure 7: Examples of synthetic face sketches on the CUFS dataset and the CUFSF dataset. From top to bottom, the examples are selected from the CUHK student database, the AR database, the XM2VTS database and the CUFSF database, sequentially.

Figure 8: Examples of synthetic face sketches on face photos in the wild.

has stronger robustness to the environment noises but tends to generate over smooth results, and the texture distribution of its synthetic sketches is inconsistent with that of training sketches. The proposed Sketch-Transformer can generate the most sketch-like textures while maintain the global structures.

Figure 8 presents some synthetic face sketches from different methods on face photos with deformations and illumination variations. Results of DGFL are able to preserve desirable structures but lose texture details. Results of pix2pix, PS2MAN tend to lose structural information and mistake the shaded area as the hair area. CycleGAN can preserve considerable structures but its synthetic sketches are more like face photos. Wild has desirable visual performance but the texture distribution of its synthetic sketches is inconsistent with that of training sketches. Our results can preserve enough structural information while generate satisfactory textures.

Table 3 presents the evaluation scores of the synthetic face photos/sketches on the CUFS dataset and the CUFSF dataset. The proposed model obtains the best score in most cases, which indicate that it achieves the best performance.

Synthetic Image	Metrics	DGFL	FCN	pix2pix	CycleGAN	PS2MAN	Wild	SCAGAN	Sketch-Transformer
CUFS Photo	LPIPS(alex) ↓	-	-	0.1993	0.2096	0.2464	-	0.1727	**0.1538**
	LPIPS(squeeze) ↓	-	-	0.1830	0.2094	0.2158		0.1643	**0.1310**
	LPIPS(vgg) ↓	-	-	0.3525	0.3882	0.3254	-	0.3053	**0.2738**
	FSIM ↑	-	-	0.7726	0.7450	0.7819	-	**0.7937**	0.7851
	FID ↓	-	-	73.56	80.44	65.04	-	80.53	**27.88**
CUFS Sketch	LPIPS(alex) ↓	0.3316	0.4517	0.2263	0.2139	0.2961	0.2807	0.2408	**0.1807**
	LPIPS(squeeze) ↓	0.2635	0.3596	0.1552	0.1529	0.2265	0.2210	0.1722	**0.1233**
	LPIPS(vgg) ↓	0.3654	0.4350	0.3734	0.3598	0.3707	0.3639	0.3627	**0.3019**
	FSIM ↑	0.7079	0.6936	0.7363	0.7219	0.7230	0.7114	0.7086	**0.7350**
	FID ↓	70.81	69.93	44.91	23.76	48.95	59.26	38.61	**20.92**
CUFSF Photo	LPIPS(alex) ↓	-	-	0.2463	0.2557	0.3145	-	**0.1735**	0.2199
	LPIPS(squeeze) ↓	-	-	0.2005	0.2002	0.2853	-	**0.1469**	0.1714
	LPIPS(vgg) ↓	-	-	0.4019	0.3791	0.4237	-	**0.3128**	0.3474
	FSIM ↑	-	-	0.7777	0.7645	0.7812	-	**0.8395**	0.7861
	FID ↓	-	-	39.82	**14.46**	78.03	-	18.84	15.22
CUFSF Sketch	LPIPS(alex) ↓	0.3524	0.4793	0.2408	0.2371	0.3288	0.3288	0.2188	**0.1971**
	LPIPS(squeeze) ↓	0.2794	0.3895	0.1628	0.1589	0.2397	0.2473	0.1500	**0.1349**
	LPIPS(vgg) ↓	0.3972	0.5305	0.3824	0.3744	0.4170	0.4053	0.3536	**0.3400**
	FSIM ↑	0.6957	0.6624	**0.7283**	0.7088	0.7233	0.6821	0.7270	0.7259
	FID ↓	57.33	124.40	35.52	14.62	64.42	59.76	18.32	**9.39**

Table 3: Quantitative results of the comparison with state-of-the-art methods on synthetic face photos/sketches of the CUFS database and CUFSF database.

5 Conclusion

In this paper, we investigate the application potential of transformer architecture (especially the self-attention mechanism) on face photo-sketch synthesis task. For this purpose, we propose a Sketch-Transformer network which consists of three closely-related modules: a MFP-Encoder, a self-attention module, and a MSPADE-Decoder. We compare the proposed models with recent state-of-the-art methods on two public datasets and face images in real scenarios. Both qualitative and quantitative results demonstrate that the proposed method achieves significant improvements in both retaining structural information and generating appropriate textures. In the future, we intend to further investigate the method of applying the self-attention module to multi-scale feature embeddings.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103202; in part by the National Natural Science Foundation of China under Grants 62036007, 61922066, 61876142, 61772402, and 62050175; in part bt the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2021JQ-198; in part by the Fundamental Research Funds for the Central Universities under Grant XJS210102.

References

[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.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

[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 *NeurIPS*, pages 2672–2680, 2014.

[Han *et al.*, 2020] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, et al. A survey on visual transformer. *arXiv:2012.12556*, 2020.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Conference and Workshop on NeurIPS*, 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.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Fei-Fei Li. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[Li *et al.*, 2016] Zhifeng Li, Dihong Gong, Qiang Li, Dacheng Tao, and Xuelong Li. Mutual component analysis for heterogeneous face recognition. *ACM TIST*, 7(3):1–23, 2016.

[Li *et al.*, 2019] Boyi Li, Felix Wu, Kilian Q Weinberger, and Serge Belongie. Positional normalization. In *NeurIPS*, pages 1622–1634, 2019.

[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.

[Mao *et al.*, 2017] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017.

[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.

[Park *et al.*, 2019] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346, 2019.

[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.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[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.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[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*, pages 83–90, 2018.

[Yu *et al.*, 2018] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Song. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *arXiv:1808.00897v1*, 2018.

[Yu *et al.*, 2020] Jun Yu, Xingxin Xu, Fei Gao, Shengjie Shi, Meng Wang, Dacheng Tao, and Qingming Huang. Toward realistic face photo-sketch synthesis via composition-aided gans. *TC*, 2020.

[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 *ICMR*, pages 627–634, 2015.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[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 *ICCV*, 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.