MOST-Net: A Memory Oriented Style Transfer Network for Face Sketch Synthesis

Fan Ji1, Muyi Sun1, Xingqun Qi2, Qi Li1, Zhenan Sun1, Excel

1CRIPAC & NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100190
2School of AI/Auto, Beijing University of Posts and Telecommunications, Beijing, China, 100876
Email: {fan.ji, muyi.sun}@cripac.ia.ac.cn, xingqunqi@bupt.edu.cn, {qli, znsun}@nlpr.ia.ac.cn

Abstract—Face sketch synthesis has been widely used in multimedia entertainment and law enforcement. Despite the recent developments in deep neural networks, accurate and realistic face sketch synthesis is still a challenging task due to the diversity and complexity of human faces. Current image-to-image translation-based face sketch synthesis frequently encounters over-fitting problems when it comes to small-scale datasets. To tackle this problem, we present an end-to-end Memory Oriented Style Transfer Network (MOST-Net) for face sketch synthesis which can produce high-fidelity sketches with limited data. Specifically, an external self-supervised dynamic memory module is introduced to capture the domain alignment knowledge in the long term. In this way, our proposed model could obtain the domain-transfer ability by establishing the durable relationship between faces and corresponding sketches on the feature level. Furthermore, we design a novel Memory Refinement Loss (MR Loss) for feature alignment in the memory module, which enhances the accuracy of memory slots in an unsupervised manner. Extensive experiments on the CUFS and the CUFSF datasets show that our MOST-Net achieves state-of-the-art performance, especially in terms of the Structural Similarity Index (SSIM).

I. INTRODUCTION

Face sketch synthesis refers to generating sketches from face photos, which attracts extensive attention in recent years. Face sketch synthesis plays an important role in a wide range of applications including digital entertainment, law enforcement, and face de-identification [1]. For instance, people tend to use sketches instead of real photos as an avatar on social media to protect their privacy. Nevertheless, it will take a vast time and effort of a professional artist to draw a dedicated sketch. The automatic sketch synthesis techniques are highly necessary. Early face sketch synthesis methods can be roughly divided into two categories: exemplar-based and regression-based methods. Exemplar-based methods [2]–[4] generate sketches by mining the relationship between corresponding input image patches and sketch patches of face-sketch pairs in the training set. However, the synthesized sketches are limited to holding the fidelity and identifiability of the face photo. Regression-based approaches formulate a linear mapping between photos and sketches as a regression problem [5], [6]. However, the mapping function is not comprehensive and accurate because of the limited capability of the linear regression strategies [7].

The last few years have seen significant growth in the use of Generative Adversarial Networks (GANs) [8] for image generation tasks. Face sketch synthesis has also undergone substantial and fundamental improvement. A great number of researchers employ various GAN models on this task [9]–[11]. Although desirable results are obtained, these approaches often encounter the over-fitting problem with the limited data. Take the Pix2Pix [10] model as an example, we could observe that the generator tries to remember all training data instead of learning the transferring strategy. As shown in Fig. 1, a string of letters appears in the upper part of the generated sketch as well as the real sketch. However, there should be no letters that do not exist in the input face photo. We infer that this is because real sketches are only used as the L1 loss supervision in the data-limited tasks. To make better use of the sketch knowledge to improve the generalization capacity of the network, the real sketch itself or its features could also be employed as a part of the input, not just for supervision.

To prevent the over-fitting problem, we propose a novel Memory Oriented Style Transfer Network for Face Sketch Synthesis (MOST-Net). Firstly, a face encoder and a sketch
encoder are utilized to obtain the face feature maps and the sketch feature maps of the same size. Secondly, inspired by [12], we construct an external memory block to store two different types of information: key memory and value memory. The key memory K stores information about the face features, which are used to compute the cosine similarity with input queries. The value memory V stores information about the sketch features, which are later used as the condition for the sketch synthesis network. This memory module can be described as a reference dictionary to find the corresponding sketch features. Meanwhile, we proposed a novel Memory Refinement Loss (MR Loss) for memory alignment with unsupervised learning. Thirdly, we inject the sketch features as style information into the generation pipeline of MOST-Net. As a result, we recombine the sketch feature slots in the memory through the queries of image features for the final sketch synthesis, which could overcome the over-fitting problem caused by singly end-to-end supervised training. And in the inference time, we only input the face images as queries and employ the memory to provide sketch-related features for the feature-level fusion. Consequently, our method achieves state-of-the-art performance in a variety of metrics on the CUFS and the CUFSF datasets.

The contributions of this paper can be summarized as:

- To the best of our knowledge, the proposed MOST-Net is the first face sketch synthesis network augmented by external neural memory networks.
- A novel Memory Refinement Loss (MR Loss) is proposed for feature alignment in the memory module. The MR Loss ensures that the key value pair of the memory network can be updated correctly according to the spatial information.
- Extensive experimental results on CUFS and CUFSF datasets show the superiority of our method.

II. RELATED WORK

A. Deep Learning-Based Face Photo-Sketch Synthesis

Deep learning-based face photo-sketch synthesis has developed rapidly in recent years. Zhang et al. [9] pioneered an end-to-end fully convolutional neural network to generate sketch. Limited by the shallow network architecture and pixel-level loss, its performance is not high-fidelity. Isola et al. [10] proposed a conditional GAN (cGAN), called Pix2pix, for general image-to-image translation tasks. Coincidentally, another profound method CycleGAN [13] enforced image to image translation by introducing a cycle consistency loss. Both of these two general backbone networks can be applied to fake sketch synthesis. Following the Pix2pix model and CycleGAN, researchers have made extensive progress. On the one hand, several works focused on improving the capability of the GAN models. Wang et al. [14] proposed multi-scale generator and discriminator architectures for synthesizing sketches. However, undesirable artifacts and distorted structures still exist in many scenarios. On the other hand, many researchers found that introducing prior information can effectively improve the performance of the sketch synthesis methods. SCAGAN [11] introduced face parsing layouts as input of cGAN to achieve state-of-the-art performance. Qi et al. [15] decomposed face parsing layouts into multiple compositions as the semantic-level spatial prior for style injection. Nevertheless, besides the model becoming more complex, they could not guarantee that the prior information is always available and beneficial.

B. Memory Network

The external memory module is used to store critical information over long periods. As a typical recurrent neural network (RNN), long short-term memory (LSTM) [16], [17] is the most classical neural network with a memory mechanism, which dominated the field of sequential data processing. Weston et al. [18] described the memory network as a dynamic knowledge base of question answering. Thus, knowledge is compressed into dense vectors to improve the neural network performing memorization. Recent research has shown that the memory network can be applied to various tasks, not only in the field of natural language processing. Yoo et al. [12] presented a memory-augmented colorization model to produce high-quality colorization with limited data. Huang et al. [19] performed image de-raining through memory-oriented semi-supervised method. Inspired by previous research, we propose a self-supervised memory mechanism for face sketch synthesis.

C. Image Style Transfer

Initial style transfer aims to transform photos into painting-like images, which indicates that face sketch synthesis can be regarded as a style transfer task. Gatys et al. [20] employed the VGG network [21] in separating image content from style, then the semantic content can be re-rendered in different styles. Ulyanov et al. [22], [23] found that Instance Normalization (IN) is far more effective than Batch Normalization (BN) in the field of style transfer. According to his research, the following studies on style transfer tends to be closely related to IN. Consecutively, Huang et al. [24] proposed Adaptive Instance Normalization (AdaIN) for real-time arbitrary image style transfer. In the meanwhile, NVIDIA proposed spatially-adaptive normalization (SPADE) [25] which injects style information into the given layout to obtain photorealistic images. The SPADE has been applied to various image synthesis tasks, including face image editing [26], face reenactment [27], and person image generation [28]. In this paper, we also borrow ideas from SPADE for sketch synthesis.

III. METHOD

Theoretically, SPADE [25] is an ideal image domain transfer method. However, in terms of the face sketch synthesis task, there is no such aligned domain-variant image as layout during the inference and test phase. Therefore, our memory-oriented network aims to construct a Memory Dictionary that maps images from photo domain \( P \) to sketch domain \( S \) on the feature level over a long period of time. Through this memory dictionary, we can still obtain this domain-variant
layout information without any input sketch in the testing. On the other hand, previous research synthesized sketches directly with the input of only real photos, in which sketches serve as supervision. Our method is able to take both photos and sketches as inputs of our end-to-end network in the training phase, which could take full advantage of ground-truth images. Fig. 2 illustrates the overall architecture of the proposed Memory Oriented Style Transfer Network (MOST-Net) for face sketch synthesis. In the following, a detailed description of proposed method is given.

A. Network Architectures

As shown in the Fig. 2, the proposed MOST-Net consists of a Photo Encoder $PE$, a Sketch Encoder $SE$, a Memory dictionary $M$, three Style Injection (SI) resblocks, and a Decoder $D$.

Given paired photo-sketch samples shown in Fig. 1 $\{(x_i, y_i) \mid x \in \mathbb{R}^{3 \times H \times W}, y \in \mathbb{R}^{1 \times H \times W}, i = 1, \ldots, N\}$, SE and PE firstly extract photo features $F_p \in \mathbb{R}^{c \times H \times W}$ and sketch features $F_s \in \mathbb{R}^{c \times H \times W}$, where $x_i$ represents the photo, $y_i$ represents the sketch, and $N$ represents the total number of samples. The SE and PE share the same structure which consists of a convolution layer and a stack of residual blocks. The memory dictionary is located after the encoders. According to Fig. 2, there are three main data flows painted with different colors. The data flow in red indicates that the memory dictionary $M$ is updated in a self-supervised manner when new training samples are fed in. The blue one shows that $F_p$ serves as a query to retrieve the most relevant slots in $M$.

Detailed updating and retrieving processes are presented in Sec. III-B. Next, we design three SI Resblocks, which inject $F_p$ as domain style information into $F_s$ to produce fake sketch representation $z(\hat{y})$. As shown in Fig. 2 ($M_1$) and ($M_2$), a SI Resblock takes $F_p$ and $F_s$ as input, which are then processed through stacks of spatial normalization, activation, and convolution operations. Motivated by [25], our method considers $F_s$ as style semantic information for realistic sketch synthesis. Finally, the decoder $D$ predicts fake sketches from $z(\hat{y})$, which consists of a stack of residual blocks followed by a convolutional output layer. Additionally, a Discriminator is employed in the same setting in Pix2Pix [10].

B. Memory Dictionary

We define the memory mechanism as a dictionary $M$ holding pairs of the key $k_i \in \mathbb{R}^c$ and value $v_i \in \mathbb{R}^c$, which represents the photo feature and its corresponding sketch feature. Each entry $(k_i, v_i)$ of $M$ represents a feature mapping from photo domain to sketch domain. This mapping function indicates the domain transferring strategy on the feature level. We consider the value corresponding to the key closest to each input photo feature to be the expression of this photo feature in the sketch domain. In this way, during model training, we
simultaneously input pairs of features to optimize the memory dictionary. In the generation stage, the photo feature is used as the query to obtain the sketch feature.

Therefore, a dictionary \( M \) containing \( K \) items can be described in the following:

\[
M = \{(k_i, v_i) \mid k_i, v_i \in \mathbb{R}^c, i = 1, 2, 3 \cdots, K\}
\]  

(1)

Additionally, \( M \) is initialized randomly, because our method do not rely on any prior knowledge, such as face parsing and salience detection.

According to Sec.III-A, the photo encoder and the sketch encoder convert input photos and sketches to photo features and sketch features. Given photo features \( F_p \) and sketch features \( F_s \), we divide them into a number of slots. For example, a face feature slot is the basic unit of complete face feature maps, which serves as a query to dictionary \( M \). Specifically,

\[
F_p = \{f_1, f_2, f_3, \cdots, f_N\}, f_i \in \mathbb{R}^c
\]  

(2)

\[
F_s = \{s_1, s_2, s_3, \cdots, s_N\}, s_i \in \mathbb{R}^c
\]  

(3)

\[N = h \times w\]

(4)

where \( f_i \) represents a face feature slot, \( s_i \) represents a sketch feature slot, \( N \) is the number of slots (feature map size), \( c \) denotes the dimension of the each slot (the number of the feature map channels). Since these slots are intended to interact with Memory Dictionary for reading and updating, they have the same dimension with keys and values.

Our memory \( M \) is updated after a photo-sketch pair is introduced to the networks in the training stage. We employ \( F_p \) to update the keys and \( F_s \) to update the values in \( M \) respectively using the same strategy. For each face feature slot \( f_i \), we firstly find the most relevant key \( k_{\text{max}} \) in \( M \) through cosine similarity which is illustrated as

\[
k_{\text{max}} = \arg \max_j (\frac{f_i \cdot k_j}{\|f_i\| \|k_j\|})
\]  

(5)

Then we update the memory key \( k_{\text{max}} \) by

\[
k_{\text{max}} \leftarrow \alpha \cdot k_{\text{max}} + (1 - \alpha) \cdot f_i
\]  

(6)

where \( \alpha \in [0, 1] \) is a decay rate. Also, each sketch feature slot \( s_i \) is used to update the values of \( M \) with the same strategy as above. During each iteration, no additional labels are required. Though we do not assign face feature slots and sketch feature slots to update specific keys and values, a key-value pair \((k_i, v_i)\) in \( M \) still represents the mapping from the photo feature slot to the corresponding sketch feature slot. Therefore, the proposed memory dictionary is trained in a self-supervised manner, which is also called as self-supervised dynamic updating in Fig. 2. Furthermore, to better align key-value pairs \((k_i, v_i)\) and to stabilize the training process, we also propose a novel Memory Refinement loss, of which detailed information is presented in Sec. III-C4.

After updating \( M \), we retrieve synthesised sketch feature \( \hat{s}_i \) by iterating the memory items through the query \( f_i \). Instead of selecting the matching item \((k_{\text{max}}, v_{\text{max}})\) to \( f_i \), we use an attentive reading strategy for obtaining \( \hat{s}_i \). We compute a weight \( a_{ij} \) for each key \( k_j \) in the \( M \), and use it to calculate a weighted average for corresponding \( v_j \) as the \( \hat{s}_i \).

\[
\hat{s}_i = \sum_{j=1}^{K} a_{ij} \cdot v_j, \quad a_{ij} = \frac{\exp(\frac{f_i \cdot k_j}{\|f_i\| \|k_j\|})}{\sum_{j=1}^{K} \exp(\frac{f_i \cdot k_j}{\|f_i\| \|k_j\|})}
\]  

(7)

Finally, we aggregate all the retrieved slots to formulate final sketch representation

\[
\hat{F}_s = \{\hat{s}_1, \hat{s}_2, \hat{s}_3, \cdots, \hat{s}_N\}, \hat{s}_i \in \mathbb{R}^c
\]  

(8)

Note that during memory reading, \( M \) is not updated using back-propagated gradients.

C. Objective Functions

The overall objective of the proposed model includes five loss functions: \( \mathcal{L}_{\text{ADV}} \), \( \mathcal{L}_{\text{rec}} \), \( \mathcal{L}_{\text{content}} \), \( \mathcal{L}_{\text{style}} \), and \( \mathcal{L}_{\text{MR}} \). In this section, we present these loss functions in detail.

1) Adversarial learning: The adversarial loss is leveraged to correctly distinguish the real sketches or generated sketches. As mentioned in Sec. III-A, following [10], the adversarial loss is formulated as:

\[
\mathcal{L}_{\text{ADV}} = E[\log D(X, Y)] + E[\log (1 - D(X, G(X))]
\]  

(9)

where \( X, Y \) denote the source photos, target sketches, \( G, D \) denote the generator and discriminator.

2) Image Reconstruction: As an image-to-image translation task, we expect that the generated sketch sketch to be as close as possible to the ground truth. Therefore, L1 distance is employed to represent the reconstruction loss, defined as

\[
\mathcal{L}_{\text{rec}} = E[\|Y - G(X)\|_1]
\]  

(10)
3) Content and Style Consistency: In order to ensure that the synthesized sketch and the target sketch have homogeneous style while synthesized sketch and the input photo have similar content, we employ pre-trained VGG-19 [21] to extract multi-level representations. In terms of fake sketch \( G(X) \) and target sketch \( Y \), we compute the L2 distance of features from VGG-19 after pool1 and pool2 as Style Loss:

\[
L_{\text{style}} = E[\sum_{i=1}^{2}||\varphi_i(Y) - \varphi_i(G(X))||_2^2]
\]  

(11) 

where \( \varphi_i(\cdot) \) represents the output features of VGG-19 net and \( i \) denotes the \( i \)th pool layer of VGG-19 net. Additionally, we compute L2 distance of \( X \) and \( G(x) \) from VGG-19 after pool4 as the Content Loss:

\[
L_{\text{content}} = E[||\varphi^4(X) - \varphi^4(G(X))||_2^2]
\]  

(12) 

The style loss and content loss make training procedure more stable.

4) Memory Refinement: With the self-supervised learning strategy, MOST-Net can realise the face sketch synthesis to a certain extent. However, the training process of \( M \) tends to be unstable and time-consuming. Thus, we design a novel Memory Refinement Loss (MR Loss) for feature alignment in the memory module to help training. As shown in Fig. 3, we define that with the same subscript \( i \), \( f_i \) and \( s_i \) should update one key-value pair simultaneously. Therefore, the most relevant key \( k_i \) of \( f_i \) and the most relevant value \( v_j \) of \( s_i \) also share same subscript \( j \). Formally, the similarity permutation matrix could formulated as:

\[
SIM_{i,j}(f_i, k_j) = \begin{cases} 
1, & k_j = k_{\text{max}} \\
0, & k_j \neq k_{\text{max}} 
\end{cases}
\]  

(13) 

as well as \( SIM_{i,j}(s_i, v_j) \). Then, we utilize the normalized L1 distance to represent MR Loss:

\[
L_{\text{MR}} = E[|SIM(f_i, k_j) - SIM(s_i, v_j)|]
\]  

(14) 

5) Full Objective: The overall loss for training MOST-Net is:

\[
L_{\text{total}} = \lambda_1 L_{\text{ADV}} + \lambda_2 L_{\text{rec}} + \lambda_3 L_{\text{style}} + \lambda_4 L_{\text{content}} + \lambda_5 L_{\text{MR}}
\]  

(15) 

where \( \lambda_1, \lambda_2, \lambda_3, \lambda_4, \) and \( \lambda_5 \) are loss weights. Besides, the generator \( G \) and the discriminator \( D \) could be optimized by:

\[
\underset{G}{\text{min}} \underset{D}{\text{max}} L_{\text{total}}
\]  

(16) 

IV. EXPERIMENTS

A. Implements Details

Our network is trained from scratch without any prior knowledge through an end-to-end training strategy. The Adam optimizer is adopted with \( \beta_1 = 0.9 \) and \( \beta_2 = 0.999 \). The learning rates are set to 0.001 and 0.00004 for the discriminator and generator respectively. For the hyper-parameters in Eq. 15, we set \( \lambda_1, \lambda_2, \lambda_3, \lambda_4, \) and \( \lambda_5 \) to 1, 200, 40, 40, and 10. Additionally, the decay rate in Eq. 6 is set to 0.999. Since we need a small decay rate for stabilizing the memory updating procedure, it takes more than 1000 epochs with a batch size of 16 for our network to converge.

B. Datasets and Evaluation Metrics

We conduct experiments on the CUFS dataset [2] and the CUFSF dataset [33]. In the CUFS dataset, there are 606 faces, of which 188 faces from the CUHK student database, 123 faces from the AR database, and 295 faces from the XM2VTS database. For each paired sample, the face photo is taken in natural lightning conditions while the sketch is drawn by the artist. The CUFSF dataset contains 1194 face-sketches pairs. However, all photos are obtained under illumination variations as illustrated in Fig. 1. For both datasets, the whole images are cropped to 200 \( \times \) 250. Thus, widely used reshaping and padding strategies [11] for face sketch synthesis are adopted to expand the input image size to 256 \( \times \) 256.

The performance of sketch synthesis is measured by multiple metrics. We first employ the Feature Similarity Index Metric (FSIM) [34] and Structural Similarity Index Metric (SSIM) [35] to evaluate the quality of synthesized sketches. Furthermore, for assessing the similarity between the ground-truth and synthesized sketch from human perspectives, we choose the Learned Perceptual Image Patch Similarity (LPIPS) [36]. In this paper, three different perceptual similarity models are used as the baseline of LPIPS, including SqueezeNet [37], AlexNet [38], and VGGNet [21].
C. Experiment Results and Analysis

The quantitative results of MOST-Net on the CUFS and the CUFSF test sets are reported in Table I. The proposed method achieves the best performance on the metrics of SSIM, which suggests that the sketch generated by our method holds the highest structural similarity with the ground-truth sketch. Because we spatially align the cross-domain features, it is reasonable for sketches to have similar structures. On the CUFS dataset, we obtain the state-of-the-art performance among all face sketch synthesis methods on the indicators of FSIM, LPIPS-SqueezeNet, and LPIPS-VGG16. Additionally, our method slightly increases the previous best LPIPS-AlexNet from 0.1807 to 0.1837. On the CUFSF dataset, there is no dominant method for all perspectives. We analyze the reason is that samples of the CUFSF dataset are non-aligned and with low-fidelity. The performance of the network targeting image-to-image translation on this data set is not stable and consistent. However, our method achieves the optimal result of LPIPS-VGG16 and sub-optimal results of LPIPS-SqueezeNet and LPIPS-AlexNet. However, our method decreases the previous best SSIM from 0.7311 to 0.7167 on the CUFSF dataset. Because LPIPS assesses similarity between two images from the human perspective [36], we consider LPIPS as a more meaningful evaluation metric for face sketch synthesis task. Therefore, we can conclude that our method could integrally construct sketches from face photos with high-fidelity and rich details and achieves state-of-the-art performances in this task.

D. Ablation Study

Furthermore, we conduct an ablation study to verify the effectiveness of the proposed memory refinement loss. Because our method utilizes a similar backbone of U-Net [39] as the Pix2Pix network, we consider Pix2Pix as the baseline. Then, we remove MR Loss from the MOST-Net to investigate its value. From Fig. 4 we can observe that sketches generated by our method are more vivid than Pix2Pix. Moreover, as illustrated in Tab. II, MOST-Net without MR Loss drastically increases SSIM from 0.4983 to 0.5431, which proves the capability of the network itself. Additionally, MR Loss slightly optimizes all indicators and further improves the sketch synthesis performance.

E. Limitation and Future Work

Although our method obtains the state-of-the-art performance for the face sketch synthesis task, there is still a lot of value left to be extracted. The primary limitation of this study is that aligned training images are needed. As we can observe in Tab. I, our method is overwhelmed by the CUFSF dataset. It is natural because we update key-value pairs through spatially aligned feature vectors. An alternative way to overcome this drawback may refer to a semi-supervised or weak-supervised learning strategy, which we are going to further study.

V. Conclusion

In this paper, we propose a Memory Oriented Style Transfer Network (MOST-Net) for face sketch synthesis. Specially, we design a novel end-to-end domain transfer framework based on an external memory module. Meanwhile, we present a Memory Refinement loss for feature alignment in the memory dictionary and stabilizing the training procedure. We conduct extensive experiments on the CUFS and CUFSF datasets. The results show that our proposed MOST-Net achieves state-of-the-art performance. Additionally, our method could be adopted as a backbone network in other image domain transfer tasks. In the future, we will conduct more qualitative and quantitative experiments to verify the feasibility of our model for general purposes. A more complete version of this research will be realized in the future.

VI. Acknowledgement

This work is founded by the National Natural Science Foundation of China under Grant U1836217.
