Face Sketch Synthesis via Semantic-Driven Generative Adversarial Network

Xingqun Qi¹,²,*, Muyi Sun²,*, Weining Wang²,†, Xiaoxiao Dong²,³, Qi Li², Caifeng Shan³,⁴,†
¹School of Automation, Beijing University of Posts and Telecommunications, Beijing, China
²Center for Research on Intelligent Perception and Computing, NLPR, CASIA, Beijing, China
³Artificial Intelligence Research, CAS, Jiaozhou, Qingdao, China
⁴Shandong University of Science and Technology, Qingdao, China
xingqunqi@bupt.edu.cn, {muyi.sun, xiaoxiao.dong}@cripac.ia.ac.cn
{qli, weining.wang}@nlpr.ia.ac.cn, caifeng.shan@gmail.com

Abstract

Face sketch synthesis has made significant progress with the development of deep neural networks in these years. The delicate depiction of sketch portraits facilitates a wide range of applications like digital entertainment and law enforcement. However, accurate and realistic face sketch generation is still a challenging task due to the illumination variations and complex backgrounds in the real scenes. To tackle these challenges, we propose a novel Semantic-Driven Generative Adversarial Network (SDGAN) which embeds global structure-level style injection and local class-level knowledge re-weighting. Specifically, we conduct facial saliency detection on the input face photos to provide overall facial texture structure, which could be used as a global type of prior information. In addition, we exploit face parsing layouts as the semantic-level spatial prior to enforce globally structural style injection in the generator of SDGAN. Furthermore, to enhance the realistic effect of the details, we propose a novel Adaptive Re-weighting Loss (ARLoss) which dedicates to balance the contributions of different semantic classes. Experimentally, our extensive experiments on CUFS and CUFSF datasets show that our proposed algorithm achieves state-of-the-art performance.

1. Introduction

Face sketch synthesis is a critical and challenging task in computer vision which refers to generating sketches from face photos. Face sketch synthesis attracts plentiful attention in recent years and plays a significant role in a wide range of applications like digital entertainment and law enforcement. In digital entertainment, face sketch is one of the most fundamental and popular portrait styles. Nevertheless, it requires vast time and efforts to create realistic face sketches by professional artists. Meanwhile, in law enforcement and criminal justice cases, clearly identifiable photos of criminals are often tough to obtain. In this case, face sketches with manifest features could provide a feasible way to promote the efficiency of the police [22]. Therefore, it is especially essential to automatically synthesize facial sketches with realistic effects and preserve identity features. So far, numerous approaches have been proposed to facilitate face sketch synthesis. Generally speaking, these methods can be roughly divided into three categories: exemplar-based approaches [19], linear-regression based approaches [39] and model-based approaches [12].

Exemplar-based approaches are the mainstream to solve face sketch synthesis tasks in early studies [19, 31, 40].
They are mainly dedicated to finding the correlation between exemplar photo patches and test photo patches in the photo-sketch paired dataset [40]. And the final output sketch is directly reconstructed by blending the exemplar patches corresponding to the test photo patches. Although these approaches achieve favorable performance in the photo-sketch paired datasets, the generated sketches have obvious drawbacks. The synthesized sketches are too smooth to hold the fidelity and identifiability of the corresponding face photo. Moreover, exemplar-based approaches need massive computational cost and inference time in the patch-wise matching process. Exemplar-based approaches are highly dependent on exemplars and have defective effects on the sketches generated by low-quality photos. Linear-regression based approaches assume that there is a linear mapping between the face photos and sketches. Then the sketches are generated by modeling this mapping from photos [11, 39]. These approaches are mostly inspired by Locally Linear Embedding (LLE) [17] which promotes the early sketch generation task due to the lower computation complexity. However, the mapping function might not be accurately formulated, resulting in the poor quality of the generated sketches [11]. Recently, the image-to-image translation task has made great progress with the development of deep learning, especially with Generative Adversarial Network (GAN) [4]. A large number of researchers are motivated to employ GAN on sketch generation tasks [29, 33, 38, 27]. The results of the face sketch synthesis also achieve a new benchmark. However, these methods are often short of the realistic detailed depiction of the synthesized sketches.

Motivated by the above researches, we propose a novel Semantic-Driven Generative Adversarial Network (SDGAN) for face sketch synthesis. Firstly, we utilize the pre-trained facial saliency detection network U2-Net [15] on the input photos to obtain the prior information of overall facial texture structure. Besides, we observe that the previous GAN networks are susceptible to illumination variations and complex backgrounds. Inspired by the great development of face parsing, we exploit the face parsing layouts as the semantic-level spatial prior to enforce globally structural style injection in the generator of SDGAN. Meanwhile, we divide face photo and sketch into local semantic classes, and propose a novel Adaptive Re-weighting Loss (ARLoss) which engages to balance the contributions of different semantic classes. Consequently, our method achieves state-of-the-art performance in a variety of metrics on CUFS and CUFSF datasets.

The main contributions can be summarized as:

- We propose a novel class-level knowledge Adaptive Re-weighting Loss (ARLoss) to balance the importance of the different semantic parts.
- We conduct extensive comparative experiments on CUFS and CUFSF datasets and obtain state-of-the-art performance.

2. Related Work

2.1. Face Photo-Sketch Synthesis

Face sketch synthesis has developed rapidly in the last few decades. Tang and Wang [19] proposed a linear transformation method to construct face sketches by taking advantage of the Principal Component Analysis (PCA). This technique is the pioneering work of exemplar-based approaches. Linear-regression based approaches benefit from their low-cost computation and appear after exemplar-based approaches [31, 40]. Zhang et al. [38] used Support Vector Regression (SVR) to obtain high-frequency information for sketch refinement. Model-based approaches are the mainstream routines of facial sketch synthesis in recent years which have gradually emerged with the boost of deep neural networks. Zhang et al. [29] utilized a fully convolutional network to generate sketch. Ji et al. [33] used multi-domain adversarial methods to construct a mapping from photo-domain to sketch-domain. Zhu et al. [38] borrowed knowledge from transfer learning and proposed a lightweight network supervised by a high-performance larger network. Recently, Yu et al. [27] decomposed the face parsing layouts into multiple compositions and encoded them into cGAN for face sketch synthesis which achieved state-of-the-art performance.

2.2. Paired Image-to-Image Translation

The image-to-image translation is often formulated as pixel-wise image generation tasks like face sketch synthesis which are applied with paired images. Isala et al. [8] proposed a conditional GAN architecture to solve the image-to-image translation task with paired input and output named Pix2Pix. Due to the eminent performance of the Pix2Pix on the paired dataset, researchers have made numerous improvements based on Pix2Pix and applied them to a wide range of other fields [16, 24]. By combining Pix2Pix and residual blocks, Wang et al. [23] proposed a novel network architecture to generate high-resolution images named pix2pixHD. Moreover, Park et al. [13] introduced the semantic layouts as spatial supervision injected in the pix2pixHD for synthesizing photorealistic images. Motivated by previous researches, we exploit cGAN like Pix2Pix as our backbone network.
2.3. Image Style Transfer

Face sketch synthesis could be regarded as a branch of image style transfer. Gatys et al. [3] successfully applied pre-trained CNNs to the Image style transfer task. Furthermore, Ulyanov et al. [20, 21] optimized the style transfer process by manipulating the Batch Normalization (BN) layers and Instance Normalization (IN) layers. Dumoulin et al. [2] utilized a group of parameters to realize the transfer of various image styles. Consecutively, Huang et al. [6] proposed the adaptive instance normalization (AdaIN) layers which could perform arbitrary style transfer without training repeatedly. Recently, Park et al. [13] put forward the spatially-adaptive normalization (SPADE) layers that inject the image style from the semantic layouts to obtain photorealistic images.

3. Method

Our semantic-driven network aims to construct a mapping from photo to sketch by utilizing semantic layouts and saliency detection. Previous researches synthesize sketches directly without taking advantage of semantic information. However, the translation from face to sketch is paired, and the maintenance of semantic information is extremely significant. Our network adopts semantic layouts to guide the generation of the sketches, especially the detailed regions. Moreover, we propose a novel Adaptive Re-weighting Loss (ARLoss) which dedicates to balance the contributions of different semantic classes. Besides, we conduct facial saliency detection on the input face photos to provide overall facial texture structure.

Given paired training photo-sketch samples \( \{(x_i, y_i) \mid x_i \in X, y_i \in Y\}^N_{i=1} \), where \( x_i \) represents photo and \( y_i \) represents sketch. The purpose of face sketch synthesis is to construct a mapping from source photo domain \( X \) to target sketch domain \( Y \). As illustrated in Fig. [1], we find that there are illumination variations and complex backgrounds in the source domain resulting in severe impacts on the identity and fidelity of the generated sketch. To handle these challenging issues, we first utilize face saliency detection results as prior information to provide the overall facial texture structure. We concatenate the texture structure \( M \) and the photo \( X \) as the input to the generator. Besides, we also employ the pre-trained face parsing network to acquire semantic layouts \( S \). Then the face semantic information \( S \) is injected into our network to produce the final synthesized result. Therefore, the overall mapping can be formulated as \( \{X, M, S\} \rightarrow Y \).

3.1. Network Architecture

Fig. [2] illustrates the overall architecture of our network. Specifically, we concatenate the paired \( M \) and \( X \) as inputs which are forwarded to the network to supply the overall facial texture structure. Pix2Pix [8] is exploited as the backbone which contains 7 convolutional and downsampling...
layers in the encoder part. In order to further strengthen the conditional semantic information in the forwarding process, we also design 7 Statistics Injection (SI) ResBlocks in the decoder, which is motivated by [13]. As shown in Fig. 2(b), the SI ResBlock consists of two convolutional layers, two ReLU layers, and two Statistics Injection (SI) modules.

As shown in Fig. 2(c), each SI module takes two inputs: the forward activation features after Batch Normalization layer and semantic masks \( S \), which are obtained by pre-trained MaskGAN [10] or BiSeNet [26]. In order to prevent semantic ambiguity, we merge all facial parts into 12 classes which are closely related to sketches: two eyes, two eyebrows, two ears, glasses, upper and lower lips, inner mouth, hair, nose, skin, neck, cloth, and background. Therefore, we have \( S = \{ s^{(1)}, \cdots, s^{(c)} \} \in \mathbb{R}^{h \times w \times c} \), where \( c \in [1, 2, \cdots, 12] \), \( s^{(c)} \in [0, 1] \), \( h \) and \( w \) denote the height and width of the feature maps. In order to inject the semantic information into the SI module, we perform the convolutional operation on \( S \) to produce the modulation parameters \( \gamma \) and \( \beta \) to normalize the final output. The produced \( \gamma \) and \( \beta \) encode sufficient spatial layout information which multiplied and added to the normalized activation through an element-wise way as shown in Fig. 2(c). In fact, the modulation parameters \( \gamma \) and \( \beta \) could provide a kind of spatial supervision from \( X \) to \( Y \) through \( S \) which are robust to illumination variations and complex backgrounds. Finally, the network structure of the discriminator keeps the same settings as Pix2Pix.

### 3.2. Adaptive Re-weighting

In previous studies, researchers always impose overall supervision and constraints on the entire generated sketches, which leads to defective performances in facial details. In this paper, we propose an Adaptive Re-weighting algorithm which could trade-off the contributions of different semantic classes, especially the detailed parts. As illustrated in Fig. 3, the synthesized sketch is represented as \( F \in R^{h_f \times w_f \times c_f} \), where \( h_f \), \( w_f \) and \( c_f \) denote the height, width and channel of the sketch, respectively. Then we enforce the element-wise multiplication between \( F \) and semantic masks \( S \) to extract the Interest-Region (IR) for each semantic class. The Adaptive Re-weighting could be formulated as:

\[
\mu (c) = \frac{1}{|S(\cdot, \cdot, c)|} \sum_{i=1}^{h_f} \sum_{j=1}^{w_f} S(i, j, c) F(i, j). \tag{1}
\]

where \( |S(\cdot, \cdot, c)| \) represents the summation of pixel numbers in each IR with the same semantic class \( c \). Obviously, this strategy normalizes each IR with \( |S(\cdot, \cdot, c)| \) which could balance the contributions of different semantic classes. Besides, \( \mu(c) \) also could be considered as the mean value of all pixels in the \( c \)-th semantic category. Furthermore, we introduce the modulation variance \( \nu(c) \) to faithfully react the semantic variation of the intra-class feature distribution. Formally, the computation is listed as follows:

\[
\nu (c) = \frac{1}{|S(\cdot, \cdot, c)|} \sum_{i=1}^{h_f} \sum_{j=1}^{w_f} \{ S(i, j, c) F(i, j) - \mu (c) \}^2. \tag{2}
\]

Note that both \( \mu \) and \( \nu \) are tensors. We named \( \mu \) and \( \nu \) as Adaptive Re-weighting (AR) maps. Finally, we construct the affinity maps between the synthesized sketch \( F \) and \( \mu \) (or \( \nu \)), which could be calculated by Cosine similarity.

\[
C_1 = \frac{F \cdot \mu}{\|F\|_2 \cdot \|\mu\|_2}, \quad C_2 = \frac{F \cdot \nu}{\|F\|_2 \cdot \|\nu\|_2}. \tag{3}
\]

Therefore, the synthesized sketch and the target sketch re-weight each semantic class of IR by constructing AR maps.

### 3.3. Objective Function

The overall objective of our model includes five loss functions: \( \mathcal{L}_{GAN}, \mathcal{L}_{content}, \mathcal{L}_{AR}, \mathcal{L}_{perceptual} \) and \( \mathcal{L}_{BCE} \).

#### A) Adversarial Loss

The adversarial loss is leveraged to correctly distinguish the real sketches or generated sketches. Follows the setting of [8], the adversarial loss is formulated as:

\[
\mathcal{L}_{GAN} = E_{X,M,Y} [\log D(X, M, Y)] + E_{X,M} [\log (1 - D(X, M, G(X, M)))] \tag{4}
\]

where \( X, Y \) and \( M \) denote the source photos, target sketches and saliency detection maps.

#### B) Content Loss

In addition, we utilize the normalized \( L_1 \) distance to represent content loss since it causes less blurring than \( L_2 \) distance.

\[
\mathcal{L}_{content}(G) = E_{X,M,Y} [\|Y - G(X, M)\|_1]. \tag{5}
\]
Figure 4. Ablation studies of synthesized sketches on the CUFS dataset. From the top to bottom, the examples are selected from XM2VTS database, AR database, and CUHK database. (a) Photo, (b) Saliency detection map $M$, (c) Semantic parsing mask $S$, (d) Target sketch, (e) CycleGAN [26], (f) Pix2Pix [8], (g) SDGAN w/ $M$, (h) SDGAN w/ $M + S$, (i) SDGAN w/ $M + S + ARLoss$, (j) SDGAN w/ $M + S + ARLoss + Perceptual Loss$, (k) SDGAN w/ $M + S + ARLoss + Perceptual Loss + BCE Loss$.

C) Adaptive Re-weighting Loss. In practice, we extract the affinity maps from target sketch and synthesized sketch, respectively. These affinity maps contain comprehensive knowledge between sketch and class-wise re-weighting maps which represent the intra-class feature distribution. Furthermore, in the process of calculating AR map, the contribution of each semantic class is adaptively re-weighted. Consequently, we reinforce the supervision on these affinity maps to constrain the generated sketch matching the feature distribution of the target domain. The Adaptive Re-weighting (AR) loss is formulated as:

$$L_{AR}(C_{target}, C_{fake}) = \sum_{r=1}^{2} \sum_{c=1}^{12} \| C_{target}^r(c) - C_{fake}^r(c) \|_2^2,$$

where the $r \in [1, 2]$ denotes two types of affinity maps.

D) Perceptual Loss. In order to ensure that the generated sketch and the target sketch have more similar specificity, we employ pre-trained VGG-19 net [18] as feature extractor to obtain high-level representations. The perceptual loss engages to make training procedure more stable.

$$L_{perceptual} = \sum_{l=1}^{2} \| \omega^l(Y) - \omega^l(G(X, M)) \|_2^2.$$

where $\omega^l(\cdot)$ represents the output features of VGG-19 net and $l$ denotes the selected pool1 and pool2 layers.

E) Binary Cross-Entropy Parsing Loss. Finally, we introduce the Binary Cross-Entropy (BCE) loss to further refine the synthesized sketch in semantic level. We contrast the semantic mask of synthesized sketch and target sketch produced by pre-trained parsing network [10] [26].

$$L_{BCE} = (\mathbb{P}(Y), \mathbb{P}(G(X, M))).$$

4. Experiments

4.1. Implements Details

Our network is trained from scratch. Both the generator and discriminator are implemented on the platform PyTorch [14] with a single NVIDIA GeForce Titan X GPU. We leverage the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The total training epochs are 200, then the initial learning rate is set to 0.0002 for the first 100 epochs and decay linearly in the last 100 epochs. Additionally, we utilize the Instance Normalization [21], and set the batchsize = 1. Meanwhile, the weighting factors are set as $\alpha = 100$, $\lambda = 100$, $\delta = 1$ and $\eta = 10$.

4.2. Datasets and Evaluation Criteria

In this article, we conduct extensive experiments on the CUHK Face Sketch Dataset (CDFS) [19] and the CUHK
Table 1. Comparison of different models in the CUFS dataset and the CUFSF dataset. ▲ indicates the higher is better, ▼ indicates the lower is better. Our method reaches the optimal and sub-optimal results in the CUFS dataset and the CUFSF dataset.

| Methods / Years | Datasets | FSIM ▲ | SSIM ▲ | FID ▼ | LPIPS (SqueezeNet) ▼ | LPIPS (AlexNet) ▼ | LPIPS (VGG-16) ▼ |
|-----------------|----------|--------|--------|--------|----------------------|-------------------|-------------------|
| CycleGAN [36] (2017) | CUFS | 0.6829 | 0.7011 | 0.4638 | 0.3753 | 58.394 | 31.262 |
| Pix2Pix [8] (2017) | CUFSF | 0.7356 | 0.7284 | 0.5172 | 0.4204 | 44.272 | 30.984 |
| MDAI [33] (2018) | CUFS | 0.7275 | 0.7076 | 0.5280 | 0.3818 | / | / |
| CycleGAN [36] (2018) | CUFSF | 0.7356 | 0.7284 | 0.5172 | 0.4204 | 44.272 | 30.984 |
| KT [41] (2019) | / | / | / | / | / | / |
| Col-cGAN [32] (2019) | / | / | / | / | / | / |
| KD+ [38] (2020) | / | / | / | / | / | / |
| MSG-SARL [1] (2020) | / | / | / | / | / | / |
| SCAGAN [27] (2020) | / | / | / | / | / | / |
| SDGAN (ours) | 0.7446 | 0.7328 | 0.5360 | 0.4339 | 33.408 | 30.594 |

Figure 5. Ablation studies of synthesized sketches on the CUFSF dataset. (a) Photo, (b) Saliency detection map M, (c) Semantic parsing mask S, (d) Target sketch, (e) CycleGAN [36], (f) Pix2Pix [8], (g) SDGAN w/ M, (h) SDGAN w/ S, (i) SDGAN w/ S + ARLoss, (j) SDGAN w/ S + ARLoss + Perceptual Loss, (k) SDGAN w/ S + ARLoss + Perceptual Loss + BCE Loss.

4.3. Results and Ablation Study

A) Results on CUFS Dataset. Table 1 shows the comparison results between our network and other state-of-the-art models in the CUFS dataset. We obtain the best performance on the indicators of SSIM, FID, LPIPS (SqueezeNet), LPIPS (AlexNet), and LPIPS (VGG-16). Our method increases the previous best SSIM from 0.5288 to 0.5360 and decreases the previous best FID from 34.2 to...
5. Conclusion

In this paper, we propose a Semantic-Driven Generative Adversarial Network (SDGAN) for face sketch synthesis by utilizing saliency detection and face parsing layouts as prior information. Specifically, we employ a semantic-injection method and propose a novel Adaptive Re-weighting strategy which dedicates to balance the contributions of different semantic classes. We conduct extensive experiments on the CUFS dataset and the CUFSF dataset. Eventually, our proposed SDGAN achieves state-of-the-art performance on these two datasets. Additionally, we will conduct more experiments on the generation of faces from sketches. A more complete version of this research will be released in the future.

Acknowledgment

This work was supported in part by the Natural Science Foundation of China under Grant No. 62076240, Grant No. U1836217, Grant No. 61721004; and in part by the Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) (Grant No. 2019JZZY010119)

References

[1] S. Duan, Z. Chen, Q. J. Wu, L. Cai, and D. Lu. Multi-scale gradients self-attention residual learning for face photo-sketch transformation. IEEE TIFS, 16:1218–1230, 2020.
[2] V. Dumoulin, J. Shlens, and M. Kudlur. A learned representation for artistic style. arXiv preprint arXiv:1610.07629, 2016.
[3] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In CVPR, pages 2414–2423, 2016.
[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. arXiv preprint arXiv:1406.2661, 2014.
