Unsupervised Sketch-to-Image Translation Network for Single Category Image Synthesis

Wanyu Deng, Xiaoting Feng, Qirui Li, Huijiao Xu

School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Shaanxi 710121
E-mail: 18291978500@163.com

Abstract: Sketch-based image synthesis is a challenging cross-modal translation problem in the field of computer vision. Existing supervised learning methods either rely on edge information or resort to category labeling. However, the amount of sketch-photo pair training data is limited. Accordingly, in this paper, we propose an unsupervised two-stage synthesis model to accomplish fine-grained sketch image translation at the pixel level. More precisely, we boldly use the grayscale image as an intermediate result of the synthesis task between the geometric transformation and the color filling process. In addition, we propose a new perceptual discriminator on the unpaired datasets to better learn the geometry of sketches, while incorporating an attention mechanism that preserves the main information of sketches to handle sketch-specific abstractions and stylistic variations. The final experimental results indicate that our approach achieves more promising results, obtaining more realistic and diverse generated images with significantly better FID scores.

Key Words: Sketch-to-Image synthesis, Unsupervised image translation, Geometric discriminator, Attention mechanism

1 Introduction

The rapid spread of various touchscreen devices has provided more convenient hand sketching and handwriting conditions for a wide range of users, and the field of computer vision has witnessed significant research advances in sketch-based image processing problems, including sketch image recognition [1,2,3], fine-grained image retrieval [4,5], and cross-domain image synthesis [6]. Obviously, sketches contain a great deal of detailed information, such as the position, texture, and shape of objects, which can convey ideas in an intuitive and flexible way and communicate the user's intentions better than other methods. Therefore, this is a clear indication of the potential value of sketch image research in terms of commercial applications. Among them, sketch image synthesis can be considered as a cross-domain image translation task, which is a remarkably challenging problem that has not been well solved yet because of these several aspects: 1) Sketches are highly abstract expressions that encompass human understanding, and sketches drawn of the same objects exhibit varying degrees of complexity and abstraction due to differences in the drawing abilities of different groups. 2) Sketches and photographs belong to two different image domains. More precisely, sketches consist of sparse random black lines and white backgrounds, while photographs contain many dense colored pixels. The huge distinction between the two domains makes the image synthesis task more difficult. 3) The tedious and labor-intensive sketch data collection process and the lack of paired data limit the development of sketch-based research. Consequently, sketch image cross-domain generation is an important research problem to be solved in the field of computer vision.

In recent years, deep learning-based image processing results have been quite abundant and shine in cross-domain image synthesis tasks, among which generative adversarial networks [7] show great potential and pix2pixHD [8] extends it to generate high-resolution images. Existing supervised learning methods use paired datasets for image synthesis [9,6], mostly combining edge information as original features and using sketches trained jointly with their edge maps to reduce domain differences. Meanwhile, there is also a part of research to change the style features of synthetic images based on example images, mostly using additional encoders to extract stylistic features and fusing cues to the synthetic image through a customized attention mechanism module to change the details information such as color and texture of the synthetic image. [25,26] proposed a reference-based module (RBNet) and a cross-domain generation correlation module (CoCosNet) respectively. In order to solve the problems of cross-domain differences and diversity of the generated images, we adopt grayscale images as an intermediate process of image synthesis. We first perform the conversion between sketch and grayscale images, and then increase the diversity and fidelity of the synthesized images through the image style conversion task, which both reduces the differences in domains and changes the fine-grained features of the synthesized images, making the generated images capable of certain retrieval.

The supervised model outperforms the unsupervised model for the cross-domain image generation task, but paired data is more difficult to obtain. Accordingly, in this work, we embark on the challenging task of completing the conversion between fine-grained sketch images at the pixel level using an unsupervised two-stage sketch image synthesis model. We propose a new geometric discriminator to learn the abstract features of sketches and fuse the perceptual losses, which can reduce the image artifacts due to missing a priori information while using an attention mechanism model to suppress the redundant rough lines during sketching, enhance the robustness of the cross-domain image synthesis task, and ensure the generated images are smoother. Experimental results show that geometric discriminators and attention mechanisms can help the network optimize the model and boost the quality of the synthesized images.

*This work is supported by Science Research Plan of Shaanxi Provincial Department of Education under Grant No. 19JC036.

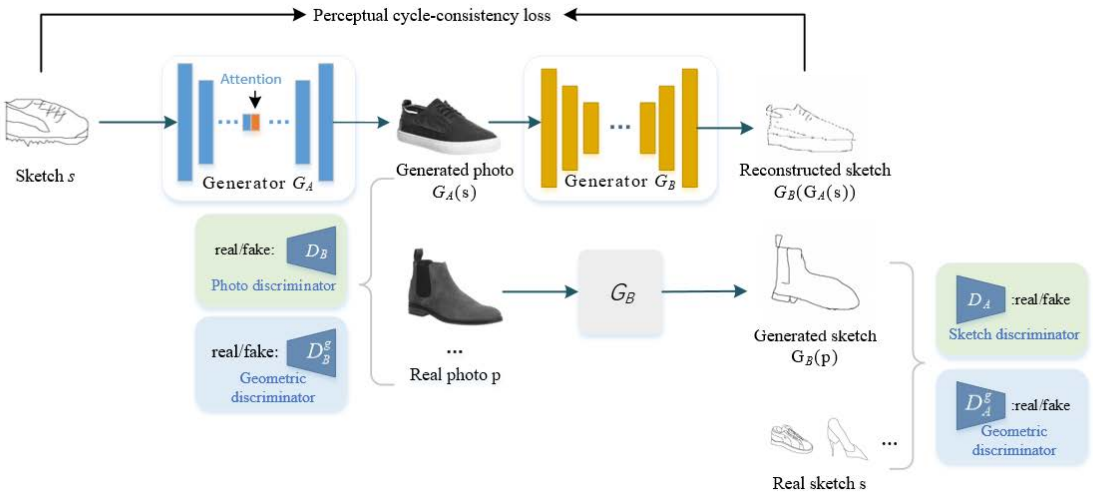

Fig. 1: Overview of the framework. It has two generators G_A: Sketch→Photo and G_B: Photo→Sketch, two discriminators D_A and D_B for the sketch and image domains respectively, and two geometric discriminators to learn the respective texture geometry features. In addition, we add an attention module to suppress or disperse the redundant rough lines.

2 Related Works

2.1 Sketch-Based Image Synthesis

The trendiest sketch-based image retrieval methods often incorporate image synthesis techniques. In the early years, classical feature descriptors were often used to construct cross-domain invariant features, such as bag-of-words representation and edge detection. Previous approaches, however, have been unable to synthesize and retrieve fine-grained image datasets. In recent years, deep learning networks have solved many complex problems in traditional image processing tasks. SketchyGAN [9] was the first attempt to generate images from multi-category sketches, decomposing the generation task into sketch complementation and recognition tasks, which prefer to supplement sketch detail information through network training and recognize it later; ContextualGAN [6] treats the sketch-based image generation problem as an image complementation problem to solve. However, none of the above studies synthesize images based on exemplary styles; Sketch2art [10] solves the style consistency of reference images but requires an additional encoder to extract style features. Liu [11] uses a self-supervised auto-encoder to separate the content and style features of sketches and RGB images. These creative artifacts have inspired research on sketch-based cross-domain image synthesis tasks and have been actively used in various practical applications such as sketch-based image editing [12,13], and 3D-sketch reconstruction [14].

2.2 Unpaired Image-to-Image Translation

Image synthesis can be viewed as an image translation task, and most algorithms concerning image translation rely on supervised settings, such as pix2pix [15]. However, paired data is hardly accessible and laborious to produce, and most of the existing work uses deep models represented by cycle-consistent generative adversarial networks (CycleGAN) [16] to learn the processing of unpaired sketches for image synthesis. The model is a cyclic mapping model that includes two mirror-symmetric GANs and cyclic consistency loss, and its improved versions include UNIT [17], a framework that assumes a shared latent space and requires corresponding images in both domains to be mapped to the same region. To make the output images diverse, MUNIT [18] further decomposes the feature space into a shared content space and a domain-specific style space to achieve multimodal image translation. It is similar to the DRIT [19] model, with the only difference being the use of weight sharing and a content discriminator to share the content space between two domains. However, the above approaches are limited to image datasets that contain good alignment between source and target domains. To address this issue, UGATIT [20] was recently proposed as a paradigm-based model that includes an attention model for aligning visual features from content and style. Liu [21] also proposed a two-stage generation model with shape transfer and content fill, respectively, for generating unsupervised sketches of reference images into photos in a single class, achieving better diversity and improved image quality. Based on this, we introduce perceptual loss and geometric discriminator that can make the generated images more realistic and consistent with human visual perception under unsupervised learning.

3 Methodology

In our work, in an attempt to reduce the differences between domains, we venture to use grayscale images as an intermediate process in the task of cross-domain synthesis of sketch images. In contrast to the edge map, the grayscale map retains both the semantic features of the sketch and is an essential attribute of the color image, containing details other than color. In addition, we use unsupervised image synthesis with unpaired datasets, which are processed in two stages: geometric transformation and color filling. Theoretically, first we employ n sketches $\{A_1 \ldots A_n\}$ to generate the corresponding n grayscale images $\{B_1 \ldots B_n\}$ through the network model, aiming to preserve their semantic features, and then render the results in color, filling them with texture detail features to obtain the final real color images $\{C_1 \ldots C_n\}$.

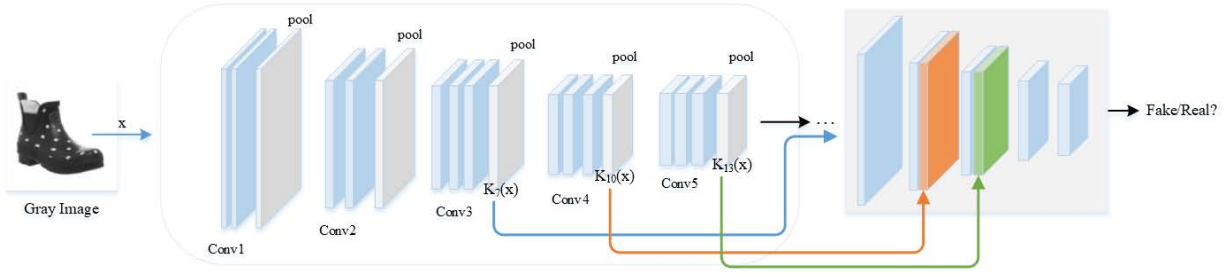

Fig. 2: Geometric discriminator

3.1 Geometric Transformation

At this stage, although our work is based on CycleGAN, which is one of the most successful unsupervised image-to-image translation frameworks, we make two improvements, as shown in Fig. 1. First, similar to [16], most of the other unsupervised methods are based on cyclic consistency constraints; it is worth noticing that reconstruction loss constraints at the pixel level can reduce the flexibility of the network and even prevent it from converging. To overcome this problem, we replace the original reconstruction loss with a perceptual cyclic consistency loss, which consists of a combination of cyclic consistency loss and perceptual loss that focuses on pixel-level information while preserving high-level features consistent with human visual perception.

Meanwhile, sketches have characteristics such as easy distortion and deformation, which have significant differences in texture and human perception. To address the sensitivity of geometric distortion to image synthesis tasks, we propose a geometric discriminator to learn the geometry of sketches and reduce the model to generate subtle local artifacts, while optimizing the generation network. The combination of PatchGAN and the geometric discriminator results in a significant enhancement in the quality of synthetic photos and the performance of human visual perception.

Fig. 1 indicates that our first phase framework consists of the following components: *two generators*: a sketch-to-grayscale generator G_A and a grayscale-to-sketch generator G_B. *two discriminators* D_A and D_B that act on each generator and are used to encourage the generators to synthesize outputs that are indistinguishable from the corresponding real image domains. On top of this, the *geometric discriminator* D^g, which is used to discriminate the geometry of sketches and optimize the generative network. Our framework is trained using unpaired sketches, defining sketch domain A and image domain B. Assuming a pair of mappings G and F, G learns the mapping from the given image a ($a \in A$) to the output image b ($b \in B$), and conversely F learns the mapping from the given image b to the output image a. The corresponding adversarial loss and cycle-consistent loss are the same as in [16].

Geometric Discriminator. In order for the CycleGAN model to learn the geometric features of sketches, we propose the geometric discriminator D^g. In contrast to PatchGAN, the input of this discriminator is a feature mapping of the loss network, acting on both real and synthetic images and sharing parameters with the perceptual cycle loss. Specifically, D^g tries to distinguish synthetic images from real images by high-level features of the sketch geometry. We perform the feature mapping on the same layer where the perceptual loss is computed. More precisely, we input the layer 7 feature map of the loss network into D^g and connect the feature map of layer 10 to the convolution layer of the first layer of D^g. Similarly, we connect the feature map of layer 13 to the output of the convolution of layer 2 of D^g. As in Fig. 2, the geometric discriminator requires five convolutional layers to output a single prediction for a given image.

Attention Module. Most human freehand sketches suffer from redundant strokes, which are rather random and discontinuous, and can interfere with the effect of cross-domain compositing of images. Accordingly, we employ the attention mechanism module to suppress or, where possible, erase redundant lines or areas in the rough sketch while retaining the major information. Assuming that the attention module generates the attention map M, which is later weighted with the feature representation of the sketch A (Eq.1):

$$T(A) = (1 - M) \odot T(A) \tag{1}$$

$T(.)$ denotes the feature map, and $\odot$ denotes the element-wise multiplication. Since the background of our selected dataset sketch is white and the lines are black, the attention mechanism makes it easy and efficient to find the areas with dense strokes.

Perceptual Loss. To generate clearer and more realistic images, we add perceptual loss to generate images that better match human visual perception. In this paper, we use a fine-tuned VGG16 network pre-trained on the Sketchy datasets [22] to extract high-level features, and use the perceptual loss for both the reconstructed and real images, combined with the cycle consistency loss to further reduce the selection space of the mapping functions G and F. The perceptual loss at layer j of the source image x and the reconstructed image y is defined as follows (Eq.2):

$$\phi^j_{per}(x, y) = \frac{1}{N_j} \| K_j(x) - K_j(y) \|_2^2 \tag{2}$$

where ϕ denotes the loss network, $K_j(.)$ denotes the layer feature map of the loss network for the input image, and N_j denotes the numbers of perceptron in layer j. We have chosen to use the same layers as the geometric discriminator, which are layers 7, 10, and 13. The average perceptual loss of the loss network is defined as follows (Eq.3):

$$\mathcal{L}^p_{cyc}(G(x)) = \frac{1}{3} E_{x \sim p_{data}} [\phi^7_{per}(x, \hat{x}) + \phi^{10}_{per}(x, \hat{x}) + \phi^{13}_{per}(x, \hat{x})] \tag{3}$$

where $\hat{x} = F(G(x))$, denotes the reconstructed image. In most cases, the combination of cyclic consistency loss, which captures low-frequency information and ensures the accuracy of low-level features of the image, and perceptual loss, which compares high-level perceptual and semantic differences between images, makes the generated images more realistic.

Ultimately, the total objective function for the training of our optimized geometric transformation phase can be expressed as:

$$\begin{aligned}\min_{G,F}\max_{D_A,D_B}\mathcal{L}(G,F) &= \mathcal{L}_{adv}(G,D_B,A,B)\\ &+\mathcal{L}_{adv}(G,D_B^g,A,B)+\mathcal{L}_{\mathrm{GAN}}(F,D_A,B,A)\\ &+\mathcal{L}_{\mathrm{GAN}}(F,D_A^g,B,A)+\lambda\mathcal{L}_{cyc}(G,F)\\ &+\mu_1\mathcal{L}_{cyc}^{p}(G(a))+\mu_2\mathcal{L}_{cyc}^{p}(F(b))\end{aligned} \quad (4)$$

where λ, μ_1 and μ_2 denote the weights of corresponding losses, respectively.

3.2 Color Filling

In the second stage, we introduce the encoder-decoder network G. The grayscale image B is used as input to obtain the final color photo C. In order to generate colorful images, we introduce the AdaIN [24] geometric transformation module to selectively inject style information into the feature map, and the resulting feature map is used as a sub-input to the decoder D for image color rendering.

Theoretically, the encoder E takes the grayscale image B and the reference image R as inputs to obtain the content feature map x = E(B) and the style feature map y = E(R), respectively. [24] adapts the channel mean and variance of the content feature map to the channel mean and variance of the style feature map by adjusting the mean and variance so that it adapts to any style transformation and then obtains a new feature map z = AdaIN(x, y) (Eq.5).

$$AdaIN(x,y) = \sigma(y)\left(\frac{x-\mu(x)}{\sigma(x)}\right)+\mu(y) \quad (5)$$

In our experiments, the reference image is optional. Content loss (Eq.6) is used to ensure that the generated image and the original image are visually perceptually identical, and style loss (Eq.7) ensures that the style information of the generated image and the reference image are spatially aligned. Where z denotes the target feature mapping obtained after encoding the content and style images through the layer, (.) denotes one of the layers in the pre-trained VGG19 model used to compute the style loss. Similar to [24], we train the decoder using the same weights of relu1_1, relu2_1, relu3_1 and relu4_1 layers to compute the style loss.

$$\mathcal{L}_c = \| E(D(z)) - z \|_2 \quad (6)$$

$$\begin{aligned}\mathcal{L}_s &= \sum_{i=1}^{k}\| \mu(\Psi_i(D(z))) - \mu(\Psi_i(R)) \|_2\\ &+\sum_{i=1}^{k}\| \sigma(\Psi_i(D(z))) - \sigma(\Psi_i(R)) \|_2\end{aligned} \quad (7)$$

Since the perceived luminance of the synthesized real image C and the input grayscale image B should be the same in the CIE Lab color space. Therefore, we add color intensity loss to train the network model (Eq.8).

$$\mathcal{L}_{color}(G) = \| B - grayscale(G(B)) \|_1 \quad (8)$$

Finally, the total objective function of our color filling phase training can be expressed as:

$$\min \mathcal{L}(G,B,R) = \eta_1\mathcal{L}_{color}(G)+\eta_2\mathcal{L}_{style}+\eta_3\mathcal{L}_{content} \quad (9)$$

where η_1, η_2, and η_3 are the weights of corresponding losses, respectively.

4 Experiments and Analysis

Dataset. We select significant fine-grained sketch image datasets for our experiments, two types of publicly available datasets: QMUL-ShoeV2 and QMUL-ChairV2 [23], with some visual texture differences. ShoesV2 contains a total of 6,648 sketches and 2,000 photos, and we used 5982/666 sketches and 1800/200 photos for training and testing, respectively. ChairV2 contains a total of 1,275 sketches and 725 pictures, and 975/300 sketches and 625/100 pictures are used for training and testing, respectively. Both contain photo instances of multiple sketches (we chose photos with at least 3 sketches for training), so the dataset is well suited to our motivation of trying to obtain modeling with diverse images. Before conducting the experiments, the original sketch images need to be uniformly sized to 256 × 256 by center cropping and normalized to make the model converge more easily.

Implementation Details. We set the training epoch for the geometric transformation phase to 480 and the color filling phase to 300. The initial learning rate is 0.0002, and we keep the learning rate constant for the first 100 training cycles, after which the learning rate decays linearly to 0. The Adam optimizer is used, and the batch size is set to 1. The rest of the parameters are set to be the same as in [20]. When training the color filling stage, we input the reference image into the network with an occurrence probability of 0.3.

Evaluation Metrics. Three different metrics are used to quantitatively evaluate the generated results: (1) **Learned perceptual image patch similarity (LPIPS)**: it measures the perceptual distance between two images and is used to evaluate the diversity of the synthesized photos. (2) **Fréchet initial distance (FID)**: which measures the feature similarity between the generated image and the real image. The image quality and diversity are evaluated based on the pool3 layer activation statistics of pre-trained InceptionV3. A lower FID value means that the generated images are less different from the real images and therefore have higher fidelity. (3) **User Preference Study (Human)**: We randomly selected 50 sets of sketch photo pairs, and the testers were required to pick one of the real photos that better matched the sketch. This is used to assess the overall perceptual sensitivity and similarity of the image, which is more reflective of real human visual perception.

The two-stage sketch-to-image training process is shown in Fig. 3. The whole training process is done on the NVIDIADGX-v1, which maintains good synthesis quality at all checkpoints while providing the required grayscale map variations. Moreover, in the color filling stage, we can

Table 1: Benchmarks on ShoeV2/ChairV2.

Model	QMUL-ShoeV2			QMUL-ChairV2		
	FID↓	Human↑	LPIPS↑	FID↓	Human↑	LPIPS↑
CycleGAN	80.69	12.0	N/A	123.48	20.0	N/A
MUNIT	94.37	10.0	**0.248**	168.81	9.5	**0.264**
DRIT	82.51	14.5	N/A	146.52	13.0	N/A
UGATIT	76.89	23.5	0.193	107.24	19.5	0.175
Proposed	**43.25**	**50.0**	0.143	**100.74**	**50.0**	0.189

Fig. 3： The grayscale image generation process with different number of iterations is shown on the ShoeV2 dataset, and the final synthesis results without(c) or with(d) the reference image. As the number of iterations increases, the appearance of the synthetic image increasingly matches human visual perception.

Fig. 4: Example results on ShoeV2 and ChairV2 datasets for the compared methods. Our model works well for the sketches and their outputs are well tuned with texture and color details in qualitative comparison.

selectively add reference images for complementary color and texture details, and the whole process demonstrates the good semantic inference capability of our model.

Comparison to baselines. To illustrate that the proposed approach yields compelling experimental results, we focused on comparing four excellent unsupervised learning methods as well as presenting the results of each qualitative comparison on the same dataset (Fig. 4). It is clear that CycleGAN[16] generates some textures, but not the general shape of the object and the results are blurred, and the overall results do not perform very well, MUNIT[18] training results preserve the geometric contours of the input sketch well, but the texture and color generation of the object is not so good, however, DRIT[19] is the opposite of the former results, preserving textures while ignoring local details differences in features. Meanwhile, UGATIT[20], the most recent work in the field of cross-domain generation, is an attention-based model that generates realistic and legible photos, but our results are more faithful to the location and contour features of the input sketches. It is also evident from the FID scores and user preference studies that our approach yields result that better match the sketches and outperforms all the baseline models currently compared. The results of the quantitative comparison of specific datasets can be found in Table 1. It is worth noting that the ChairV2 dataset is much smaller than the Shoes dataset, so our experimental results perform better on the latter data. In addition, Fig. 3 and Fig. 5 confirm the effectiveness of the different levels of features generated in the first stage for generating realistic color images.

Objectives Ablation. To explore the effect of adding modules in the middle of the network, we set up comparison experiments with different modules. The results of the comparison on qualitative evaluation are shown in Table 2, and the results clearly confirm that our added attention mechanism module and geometric discriminator module improve the semantic accuracy of the synthesized images and are effective in ensuring the quality and diversity of the generated images.

Table 2: The quantitative comparison of our framework with different architecture designs. The FID results were all calculated at the end of the complete experimental procedure.

FID↓	CycleGAN	w/o Attention	w/o D^g	Our
ShoeV2	80.96	50.42	48.29	43.25
ChairV2	123.48	107.86	103.24	100.74

Fig. 5: Sketch-based photo synthesis results. left: input sketch, right: synthesized photo. Results obtained on ShoeV2 and ChairV2.

5 Conclusion

In this work, we adopt an unsupervised two-stage learning mechanism to achieve sketch-to-photo synthesis. The qualitative and quantitative results of this paper show that the combination of a perceptual discriminator and an attention mechanism can better learn to sketch high-level features, and the model can synthesize clearly distinguishable and realistic images using unpaired data, which can preserve detailed information and increase the diversity of images while providing new ideas for some cross-domain image retrieval and recognition studies.

References

[1] Q Yu, Y Yang, F Liu, et al. Sketch-a-Net: A Deep Neural Network that Beats Humans. *International Journal of Computer Vision*, 122(3): 411-425, 2017.

[2] H Zhang, P She, Y Liu, et al. Learning Structural Representations via Dynamic Object Landmarks Discovery for Sketch Recognition and Retrieval. *IEEE Transactions on Image Processing*, PP(99): 1-1, 2019.

[3] J Qi, M Yu, F Xin, et al. Sequential Dual Deep Learning with Shape and Texture Features for Sketch Recognition. 2017.

[4] A K Bhunia, P N Chowdhury, A Sain, et al. More photos are all you need: Semi-supervised learning for fine-grained sketch-based image retrieval, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4247-4256, 2021.

[5] A K Bhunia, Y Yang, T Hospedales, et al. Sketch Less for More: On-the-Fly Fine-Grained Sketch Based Image Retrieval. *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2020.

[6] Y Lu, S Wu, Y W Tai, et al. Image generation from sketch constraint using contextual gan, *Proceedings of the European conference on computer vision*. 205-220, 2018.

[7] I Goodfellow, J Pouget-Abadie, M Mirza, et al. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[8] T C Wang, M Y Liu, J Y Zhu, et al. High-resolution image synthesis and semantic manipulation with conditional gans, *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8798-8807, 2018.

[9] W Chen, J Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9416-9425, 2018.

[10] B Liu, K Song, Y Zhu, et al. Sketch-to-art: Synthesizing stylized art images from sketches, *Proceedings of the Asian Conference on Computer Vision*. 2020.

[11] B Liu, Y Zhu, K Song. Self-Supervised Sketch-to-Image Synthesis, *Proceedings of the AAAI Conference on Artificial Intelligence*. 35(3): 2073-2081, 2021.

[12] T Portenier, Q Hu, A Szabo, et al. Faceshop: Deep sketch-based face image editing. arXiv preprint arXiv:1804.08972, 2018.

[13] J Yu, Z Lin, J Yang, et al. Free-form image inpainting with gated convolution, *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4471-4480, 2019.

[14] L Yang, J Wu, J Huo, et al. Learning 3D face reconstruction from a single sketch[J]. Graphical Models, 115: 101102, 2021.

[15] P Isola, J Y Zhu, T Zhou, et al. Image-to-image translation with conditional adversarial networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125-1134, 2017.

[16] J Y Zhu, T Park, P Isola, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks, *Proceedings of the IEEE international conference on computer vision*. 2223-2232, 2017.

[17] M Y Liu, T Breuel, J Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

[18] X Huang, M Y Liu, S Belongie, et al. Multimodal unsupervised image-to-image translation, *Proceedings of the European conference on computer vision*. 172-189, 2018.

[19] H Y Lee, H Y Tseng, J B Huang, et al. Diverse image-to-image translation via disentangled representations, *Proceedings of the European conference on computer vision*. 35-51, 2018.

[20] J Kim, M Kim, H Kang, et al. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. arXiv preprint arXiv:1907.10830, 2019.

[21] R Liu, Q Yu, S X Yu. Unsupervised sketch to photo synthesis, *European Conference on Computer Vision*. Springer, Cham, 36-52, 2020.

[22] P Sangkloy, N Burnell, C Ham, J Hays. The sketchy database: Learning to retrieve badly drawn bunnies. In: *SIGGRAPH* 2016.

[23] Q Yu, F Liu, Y Z Song, et al. Sketch me that shoe, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 799-807, 2016.

[24] X Huang, S Belongie. Arbitrary style transfer in real-time with adaptive instance normalization, *Proceedings of the IEEE international conference on computer vision*. 1501-1510, 2017.

[25] P Zhang, B Zhang, D Chen, et al. Cross-domain correspondence learning for exemplar-based image translation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5143-5153, 2020.

[26] J Lee, E Kim, Y Lee, et al. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 5801-5810, 2020.