SketchBird: Learning to Generate Bird Sketches from Text

Shaozu Yuan[1], Aijun Dai[2], Zhiling Yan[3], Zehua Guo[4], Ruixue Liu[1], and Meng Chen[*1]

[1]JD AI, China [2]HKUST, Hong Kong [3]Sun Yat-sen University, China [4]UC San Diego, USA

Abstract

Sketch plays a critical role in the human art creation process. As one of the functions of the sketch, text-to-sketch may help the artists to catch the fleeting inspirations efficiently. Different from traditional text2image tasks, sketches consist of only a set of sparse lines and depend on very strict edge information, which requires the model to understand the text descriptions accurately and control the shape and texture in the fine-grained granularity. However, there was very rare previous research on the challenging text2sketch task. In this paper, we first construct a text2sketch image dataset by modifying the prevalent CUB dataset. Then a novel Generative Adversarial Network (GAN) based model is proposed by leveraging a Conditional Layer-Instance Normalization (CLIN) module, which can fuse the image features and sentence vector effectively and guide the sketch generation process. Extensive experiments were conducted and the results show the superiority of our proposed model compared to previous baselines. An in-depth analysis was also made to illustrate the contribution of each module and the limitation of our work.

1. Introduction

Sketch is a set of human-drawn strokes imitating the approximate boundary and internal contours of an object. It plays a vital role in art composition, acting as an indispensable intermediate state. The artists firstly transfer the initial idea into a sketch with abstract content and ambiguous semantics, then complement and refine it repetitively with more elaborate details, including composing the layout, colorizing the figures and filling the texture, etc, and finally finish the artistic paintings. Generating sketches from natural language descriptions can be used as a functional approach to initialize the art composition, which may help the artist speed up the representational process and catch the flitting inspirations efficiently.

*Corresponding author. Email: chenmeng20@jd.com

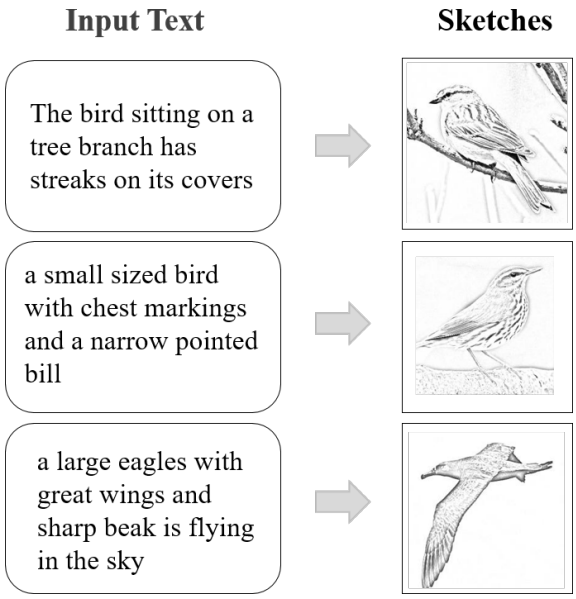

Figure 1. Text-to-sketch examples in our SketchCUB dataset.

Unlike generating realistic images conditioned on given text descriptions (aka. text2image), text-to-sketch (text2sketch) is a more challenging task. Since sketch generation drops the color feature and only keeps the stroke feature, it depends on rigid edge information. Thus the mission requires the computer to identify the critical characteristic and draw them with simple smooth strokes while understanding and getting the point from the limited natural language. Besides, any unreasonable stroke will bring a strange visual perception for human eyes. However, to the best of our knowledge, there is very rare previous work exploring this challenging task. Most researchers mainly focus on two separate tasks, image generation from text and sketch generation from image [14, 22]. Various advanced Generative Adversarial Networks (GANs) have been proposed, including Deep Fusion GAN with a fusion module to deepen the text-image fusion process [25]. Whereas on sketch generation from images, researchers

have contributed most in human facial sketch applications [14, 32, 24, 16].

To bridge the gap of text to sketch, in this paper, we try to challenge the task of generating sketches from natural language descriptions. There are several obstacles. Firstly, there is no off-the-shelf mature dataset with high-quality text and sketch pairs as far as we know. On the other hand, most traditional text2image approaches [31] are based on multi-stage modular architecture and lack an efficient fusion mechanism between image features and sentence vector, which ignore the particularity of sketches and thus can not control the amount of change in shape and texture based on the input natural language text accurately [11]. The above limitations make the existing text2image models perform less satisfying on this task.

To address the above challenges, we propose a novel text-to-sketch generation approach named text2sketch Generative Adversarial Network (T2SGAN). We follow the natural process in art generation and focus on the sketch generation from the natural language descriptions. Moreover, to facilitate the quality of generated sketches, inspired by U-GAT-IT [15], we modify the existing GAN model and implement Conditional Layer-Instance Normalization (CLIN) based fusion block in the T2SGAN's generator, which has the following three advantages. Firstly, compared to multiple generators and discriminators structure in the traditional text2image models, CLIN can fuse the image features and sentence vector seamlessly. It also encourages the text descriptions to guide the whole image generation process. Secondly, compared to previous popular batch normalization (BN), CLIN does not need to depend on the size of the batch of images and the training process becomes more stable. Thirdly, by performing the normalization in the layer and channel on the feature map based on the global sentence vector, CLIN is capable of flexibly controlling the amount of change in the generated sketches' shape and texture features with learned parameters from the dataset [15]. All the above advantages finally directly enhance the performance of T2SGAN's generator and indirectly improve the accuracy of T2SGAN's discriminator.

To tackle the deficiency of the text2sketch dataset, we contribute a new dataset (denoted as SketchCUB) originated from the prevalent text2image CUB [27] dataset. Specifically, we apply an efficient model-based approach to transfer the realistic bird images into sketch bird images. For the text descriptions, we drop the words describing the color characteristic which have no impact on the matched sketch image. Meanwhile, we delete the text without obvious meaning and rephrase some natural language descriptions without clear meaning. Figure 1 illustrates some examples from our dataset.

Overall, our contributions are summarized as follows:

- We propose a novel text2image task, which is to generate sketch images from natural language descriptions. We also contribute a dataset (denoted as SketchCUB) by modifying the prevalent CUB dataset to facilitate the future research.
- We devise a novel GAN-based model (named as T2SGAN) for this task by leveraging several Conditional Layer-Instance Normalization (CLIN) fusion blocks in the generator. Extensive experiments were designed to verify the effectiveness and superiority of our proposed model. To help other researchers better replicate our work, both our dataset and code will be released publicly.

2. Related Work

To the best of our knowledge, text2sketch has not been wildly explored in deep learning area, but there are some researches related to this subject. Blow we summarize related works in two aspects including text2image generation and image2sketch generation.

2.1. Text to Image

Generative Adversarial Networks (GANs) [6] are a framework that gives a lot of inspirations in the deep learning research field. Many variants have been explored based on GAN and were largely used in the area to generate the images from text descriptions. For instance, [18] is the first to use the conditional GAN (cGANs) to produce plausible images from text descriptions. DM-GAN [37] uses the idea of Memory Network [29] to build a dynamic memory module that can refine the blurry information which is not well generated from the initial images. DF-GAN [25] introduces a one-stage text-to-image backbone that synthesizes the image by one pair of generator and discriminator instead of using different generators. Attn-GAN [31] establishes a cross mechanism that evaluates both the text description and the image information and computes the image-text matching loss to add more details while training the generator. StackGAN [34] uses multiple pairs of generators and discriminators to generate high-resolution images based on the low-resolution images and intermediate features gathered from the text information.

2.2. Sketch Generation

Comparing to natural images, sketches are rare but useful in both research and application area. Sketch generation has been studied in different approaches. [3] introduces Edge-GAN which uses the encoder to capture the edge spatial layout of the sketches. [4, 35] introduce an automatic photo-sketch synthesis and retrieval algorithm based on the sparse representation. ASGAN [24] generates the face sketches by two pairs of generators and discriminators

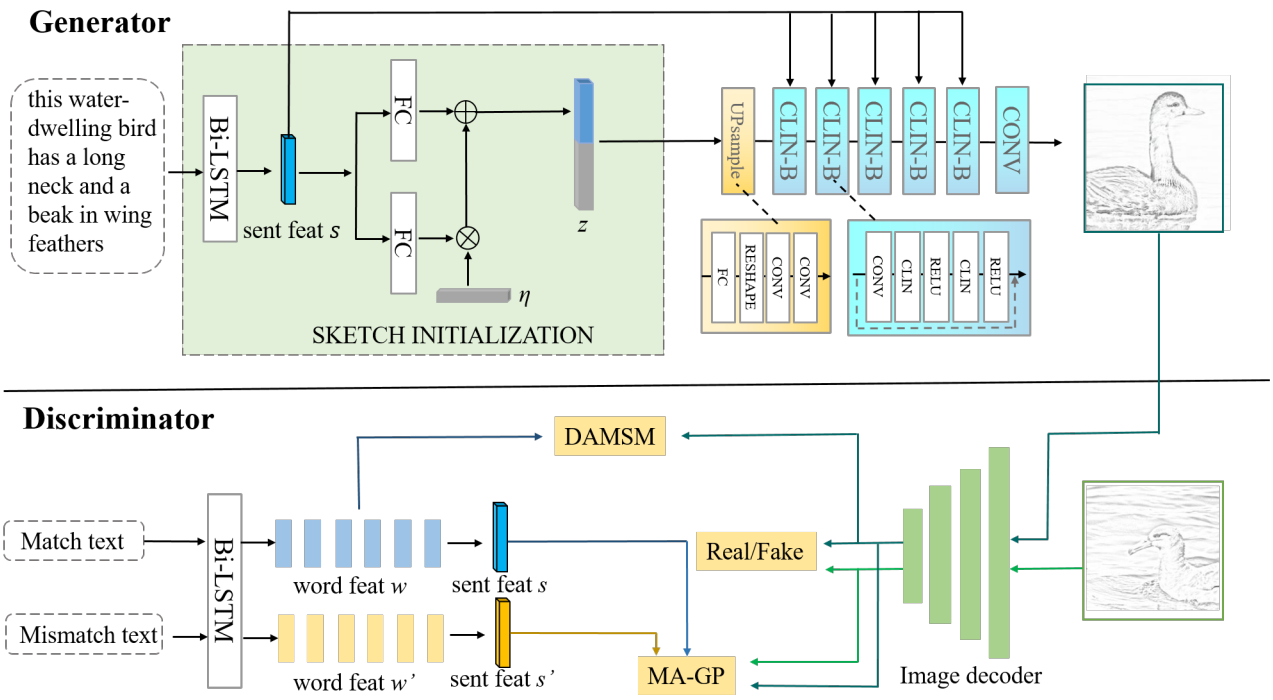

Figure 2. Overall structure of T2SGAN for text-to-sketch generation. z and η are noisy vectors for sketch initialization. s and w are matched sentence-level and word-level representations for the real sketch. s' and w' are corresponding representations of negative samples.

that one of which is to gather the attributes of faces while the other one is used for converting image to sketch. SketchGAN [17] employs a conditional GAN model to generate the missing part while comparing the free-hand sketch to the actual image. DoodlerGAN [5] proposes an idea about creating sketches that were not seen in human drawings. Finally, Stagewise-GAN [28] introduces a text2sketch model that is particularly used in the field of the criminal investigation. It encodes the face attributes described in the text and transforms that into sketches by a model that can refine the complex and ambiguous description of the text.

Our purposed method T2SGAN is different from above two lines of research including the traditional text2image task and image2sketch task. We take a short path that forms the sketch directly from the text. Comparing to the Stagewise-GAN which focuses on the attributes of the human faces, our method provides a more detailed and complex sketches based on the description of birds. Our model gathers the features of text and images more deeply and efficiently by the sequence of CLIN Blocks. Thus, our model is more effective to synthesize text-matching sketches and covers more precise details.

3. T2S-GAN

In this paper, we propose a novel text2sketch Generative Adversarial Network named as T2SGAN. The architecture of T2S-GAN is shown in Figure 2. In this section, we elaborate each part of our model as follows: First, we introduce the overall framework of T2SGAN and describe the components of the generator and discriminator. Then, we illustrate the proposed Conditional Layer-Instance Normalization (CLIN) fusion block in detail. Finally, we introduce the loss function of each part.

3.1. Overall Framework

Text encoder. The text encoder aims at learning the feature representations from the natural language descriptions. Here, the bidirectional long short-term memory (Bi-LSTM) network [7] is applied and fine-tuned by AttnGAN [31] with our dataset as text encoder. It is trained with real image-text pairs by minimizing the Deep Attentional Multimodal Similarity Model (DAMSM) loss [31]. The text encoder encodes the sentence into a sentence vector. We adopt the last hidden state vector as sentence vector s and the hidden states on each step as word vector w.

Sketch initialization. To avoid discontinuity in the latent data manifold with limited data, we follow [34] to augment the sentence vector with a random vector η. Then the augmented sentence feature s is further concatenated with a random vector z as initialized input sketch feature to the network. Here, z and η are both sampled from a standard normal distribution.

Generator. The generator has two inputs, a sentence vector from text encoder and an initialized sketch vector. The initialized sketch vector is firstly fed into a fully connected layer and the output is reshaped to sketch features. It also includes a designed up-sampling block which consists of fully connected layer, a reshape operation and two convolution layers to generate a sketch feature map. Then five CLIN fusion blocks are implemented to fuse the text and image features. Residual block is finally utilized to fuse the text information and visual feature maps to learn multi-modal representations across image and text features.

Discriminator. Following the adversarial training strategy in GANs, we employ a discriminator D_i that distinguishes whether the sketch is real or fake to guide the generation corresponding to its relevant ground truth. It is composed of several down sampling blocks and convolution layers. To be specific, to further enforce our model to learn better alignment between the image and the conditioning text, we adopt the Matching-Aware Gradient Penalty (MA-GP) [18] and design a discriminator D_m during training. The discriminator takes real sketches and their corresponding text descriptions as positive sample pairs. The negative sample pairs are composed in two ways: the first group includes real images with mismatched text, whereas the second one includes the synthetic images with their corresponding text. Here, the text is fed to text encoder to obtain text embeddings, and the sketch is fed through a series of down-sampling blocks to obtain sketch features. Finally, the sketch features and text features are concatenated to produce the decision score.

3.2. CLIN Fusion Block

Fusion of features from different modalities is the core block in many multi-modal generation tasks. AttnGAN [31] adopts the attention mechanism to extract details from the text to generate the corresponding visual concepts. SD-GAN [33] uses batch normalization (BN) to reinforce the visual-semantic embedding for the visual generation and proves its superiority in generation task. However, the advantage of BN is not obvious for sketch generation since BN considers the content of all images in a batch when calculating the normalized features. As a result, the unique details of each sketch sample is neglected. Compared with general image-to-image generation tasks, sketch generation is a much more challenging task. As a visually sensitive task, slight change of sketch line may affect the quality of the generated sketch. Unreasonable strokes in white canvas will also lead to strange visual perception for human eyes. Due to the sensitivity of this task, the information of each pixel for each sample is very important. Therefore, BN, an algorithm that normalizes all samples in each batch is not suitable for sketch generation. Inspired by AdaLIN [15], we propose Conditional Layer-Instance Normalization (CLIN) by combining the advantage of Instance Normalization (IN) and Layer Normalization (LN) to selectively change or keep the content information. By this way, CLIN fusion block is devised to facilitate the complex text2sketch problem.

$$\beta = FC(s) \tag{1}$$

$$\gamma = FC(s) \tag{2}$$

CLIN can help our attention-guided model to flexibly control the shape and texture based on the input natural-language text. The sentence vector s is normalized by two one-layer Fully Connected layer (FC) to predict the language-conditioned channel-wise scaling parameter γ and shifting parameter β respectively:

$$CLIN(x, \gamma, \beta) = \gamma(\rho LN(x) + (1-\rho)IN(x)) + \beta \tag{3}$$

$$IN = \frac{a - \eta_I}{\sqrt{{\iota_I}^2 + \zeta}} \tag{4}$$

$$LN = \frac{a - \eta_L}{\sqrt{{\iota_L}^2 + \zeta}} \tag{5}$$

where η_I, η_L, and ι_I, ι_L are channel-wise, layer-wise mean and standard deviation correspondingly. The values of ρ is adjusted by the generator in the range of [0, 1] to balance the weight of LN and IN. When the value of ρ is close to 1, it means the LN is more important, vice versa, the IN is more important when the value of ρ is close to 0.

As shown in Figure 2, the CLIN fusion block (CLIN-B) consists of a convolution layer, multiple stacked CLIN and ReLU layers. As the CLIN fusion blocks deepen the network, a residual connection[8] is equipped for each CLIN fusion block to avoid vanishing gradient during training.

3.3. Loss Function

Discriminator Loss. To promote the semantic consistency between text and the generated sketch for the discriminator, the discriminator is trained via adversarial loss associated with the MA-GP [25] loss as follows:

$$\begin{aligned}\mathcal{L}_{adv}^{D} = & -\mathbb{E}[min(0, -1 + D_m(I, s))] \\ & - 1/2\mathbb{E}[min(0, -1 - D_m(I', s))] \\ & - 1/2\mathbb{E}[min(0, -1 - D_m(I, s'))] \\ & + \mathbb{E}[(k\nabla_x D_m(I, s) + k\nabla_e D_m(I, s))^p]\end{aligned} \tag{6}$$

where s is the corresponding text description of real sketch I while s' is the mismatched text. I' is the generated sketch; k and p are two hyper-parameters to balance the effectiveness of gradient penalty.

Generator Loss. To make the generated image I' similar to real image I and makes it match the corresponding input

text, the whole adversarial loss of generator with Matching-Aware Gradient Penalty(MA-GP) is as follows:

$$\mathcal{L}_{adv}^{G} = -1/2[\mathbb{E}[D_i(I')] + \mathbb{E}[D_m(I', s)]] \quad (7)$$

We also apply the perceptual loss [13] based on a ResNet network R pre-trained on the ImageNet dataset [19], to keep the generated sketch semantically consistent with real sketch in content. The network is used to extract semantic features from both the generated image $\hat{I}$ and the real sketch I, and the perceptual loss is defined as:

$$\mathcal{L}_{per}^{G}(\hat{I}, I) = \frac{1}{W_i H_i C_i}(||R_i(I) - R_i((\hat{I}))||^2) \quad (8)$$

where R_i is the activation of the ith layer of the ResNet network, W_i,H_i and C_i represents width, height and channel of current feature map.

Until this point, there is no guarantee that the output sketch of our generator will be corresponding to the text in fine granularity. To solve this issue, we add the widely-used DAMSM proposed by [31] in our framework. DAMSM loss $\mathcal{L}_{dam}$ provides an additional word level sketch-text matching loss for training the generator.

Finally, we jointly train the generator, discriminator by using the full objective as follows:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{per} + \lambda_2 \mathcal{L}_{dam} \quad (9)$$

where λ_i controls the scale of loss weight to balance different losses.

3.4. Implementation Details

For text encoder, the embedding size of bidirectional LSTM is 256, the maximum length of the text is set to 18. And the noisy vector is initialized with dimension of 100. In the genetator, the initialized sketch vector is reshaped to 512×4×4. The up-sampling blocks consist of the nearest-neighbor upsampling followed by a 3×3 convolution with stride of 1. The CLIN block consists of convolution layers with size 3×3 and stride 1. The generated sketch is in the resolution of 256×256. Our network is trained using Adam with $\beta_1 = 0.0$ and $\beta = 0.9$. The learning rate is set to 0.0001 for the generator and 0.0004 for the discriminator. The model is trained for 500 epoches with batch size of 24. The hyper-parameter of λ_1 and λ_2 are set to 0.1 and 0.2 respectively, and p in Equation 6 is set to 5 by following [36].

4. Experiments

To validate our method, we conduct extensive quantitative and qualitative evaluations. We select two typical methods on text-to-image synthesis as baseline methods for comparision: **AttnGAN** [31] and **DM-GAN** [37]. Results by the two baseline methods are obtained by their authors' publicly released code. In addition, we design a T2S-GAN variant by using Conditional Batch Normalization (**CBN**) [33] to investigate the proposed components of our proposed CLIN. We also conduct further model tuning and error analysis to explain the hyper-parameter setting and limitation of our work.

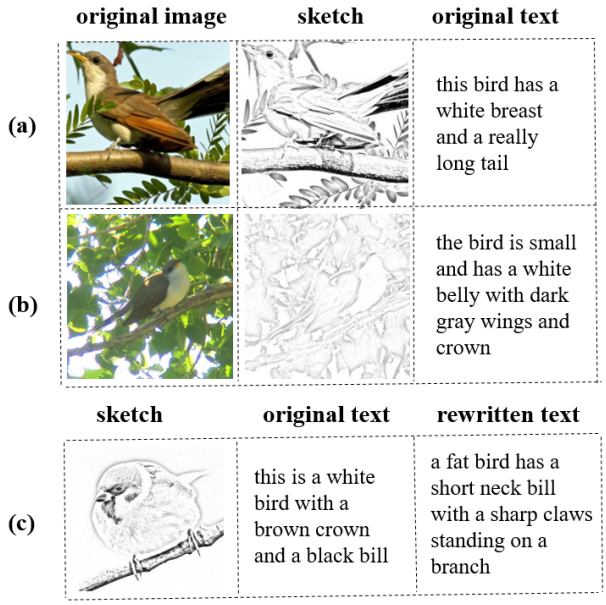

Figure 3. The construction process of SketchCUB dataset.

4.1. Datasets

As the deficiency of text-to-skech dataset, we construct a large-scale image dataset, SketchCUB, for experiments. The dataset is mainly converted from CUB dataset [27], where each bird image includes two to six descriptive sentences. The construction process contains two steps. First, we modify the natural descriptions for each image. Considering that the sketch is not colorized, we remove the color description words and rewrite the text based on the content of the image manually. Totally 30 graduate students were invited to rephrase and check the natural descriptions twice. Second, we apply an open-source holistically-nested network (HED) [30] to transfer realistic images into sketches, which will be explained later. Finally, sketches images that are not easily identified or of low quality were also removed. As shown in Table 1, the final SketchCUB dataset contains 200 bird categories with 10,843 images. It includes a training set with 8,326 images in 150 categories and a test set with 2,517 images in the remaining 50 categories.

To obtain the sketches from colored images, we utilize the holistically-nested network (HED) [30] to extract sketch. HED is proposed to address the edge detection issue, which extracts features on an object level. Recent

Table 1. Statistics of SketchCUB dataset.

Dataset	category	image
train	150	8,326
test	50	2,517
Total	200	10,843

works [1, 12] have proved the HED's capability of generating sketches. Inspired by those works, we connect a scale-associated side output to the first convolutional layer in the holistically-nested network, to extract the expected characteristics and fine-grained information. Among the fusion layer in HED, we sum up the multi-scale detection responses to prevent fine-grained features from being neglected during sketch generation. It also helps to resolve the ambiguity in boundary detection and improve the performance of edge detection [2, 21]. We manually evaluated the generated sketches from HED and verified the high quality of the images.

The process of sketch extraction is shown in Figure 3 (a) and (b). It indicates the text2sketch generation in bird species is a challenging task. On the one hand, colorized images may help to distinguish the species difference between various birds. However, sketches are characterized by abstractness, making bird pieces visually indistinguishable. On the other hand, some of the sketch images includes both foreground objects (usually the birds) and background objects (trees or branches), thus presenting specific detailed appearances for the foreground bird is more challenging. In this case, we only keep (a) and remove (b) as background in (b) is dominated and make the bird sketch not clear. Figure 3 (c) present the example of de-colorized sketch and re-written text description.

4.2. Baseline and Evaluation Metric

For baselines, we use AttnGAN and DM-GAN for performance comparsion:

- **AttnGAN** [31] consists of an attentional generative network which generates high quality image through a multi-stage process. It also proposes a DAMSM loss to compute the fine-grained image-text matching loss for training the generator. This designed structure allows AttnGAN to produce attention-driven, multi-stage refinement for fine-grained text-to-image generation.
- **DM-GAN** [37] introduces a dynamic memory module to refine fuzzy image contents. It also includes a memory writing gate to highlight important text information and a response gate to fuse image and memory representations. The proposed architecture enables DM-GAN to refine initial images with wrong color and rough shapes during text-to-image generation tasks.

The performance of these models is evaluated by Inception Score (IS) [20], Frechet Inception Distance (FID) [9], and human evaluation. Each model generates 3000 images conditioning on the captions from the test set for evaluation:

- **FID** uses the pre-trained Inception v3 network [23] to calculate the Frechet distance between synthetic and photo-realistic images based on the extracted features [9]. A lower FID implies a closer distance between synthetic image distribution and photo-realistic image distribution.
- **SSIM**[10] is an image quality evaluation algorithm to measure the similarity of two images. When two images are identical, the value of SSIM is 1. The larger the value is, the better the image quality is.

For human evaluation, we invite 15 specialists in fine arts to evaluate the synthesized sketch images among different models from both perspective of **visual quality (VQ Score)** and **semantic consistency (SC Score)** between the given descriptions and generated sketches.

4.3. Quantitative Analysis

The experimental results against baselines are reported in Table 2. As shown below, our method achieves the highest SSIM score and the lowest FID score compared with other models and the model variant with CBN blocks.

In particular, compared with AttnGAN [31] which uses a cross mechanism to combine the features of text and image, our model improves the SSIM metric from 0.41 to 0.45 (9.76% relative improvement) and decreases the FID metric from 55.55 to 26.62 (52.08% relative improvement) on the SketchCUB dataset. Compared with DM-GAN [37] which employs Memory Network to handle the information of the image, our model also improves SSIM from 0.36 to 0.45 (25% relative improvement) and decreases FID from 33.51 to 26.62 (20.56% relative improvement) on the SketchCUB dataset. Finally, when we replace the CLIN block with CBN block, both FID and SSIM metrics degrades significantly, which demonstrates the advantage of CLIN. The quantitative comparisons of SSIM and FID show that our proposed model T2S-GAN can generate higher-quality sketches from text descriptions.

Additionally, we also verify the contribution of additional perceptual loss and the DAMSM loss by conducting ablation study. It's observed that, after removing each loss, both the FID and SSIM metrics degrade, which indicates the necessity of each component.

4.4. Qualitative Analysis

We also compare and analyze the generated sketch images generated by AttnGAN, DM-GAN, and the proposed T2S-GAN. The criteria includes both the quality of the synthesized sketches and text-sketch semantic consistency. Figure 4 illustrates the generated sketch examples.

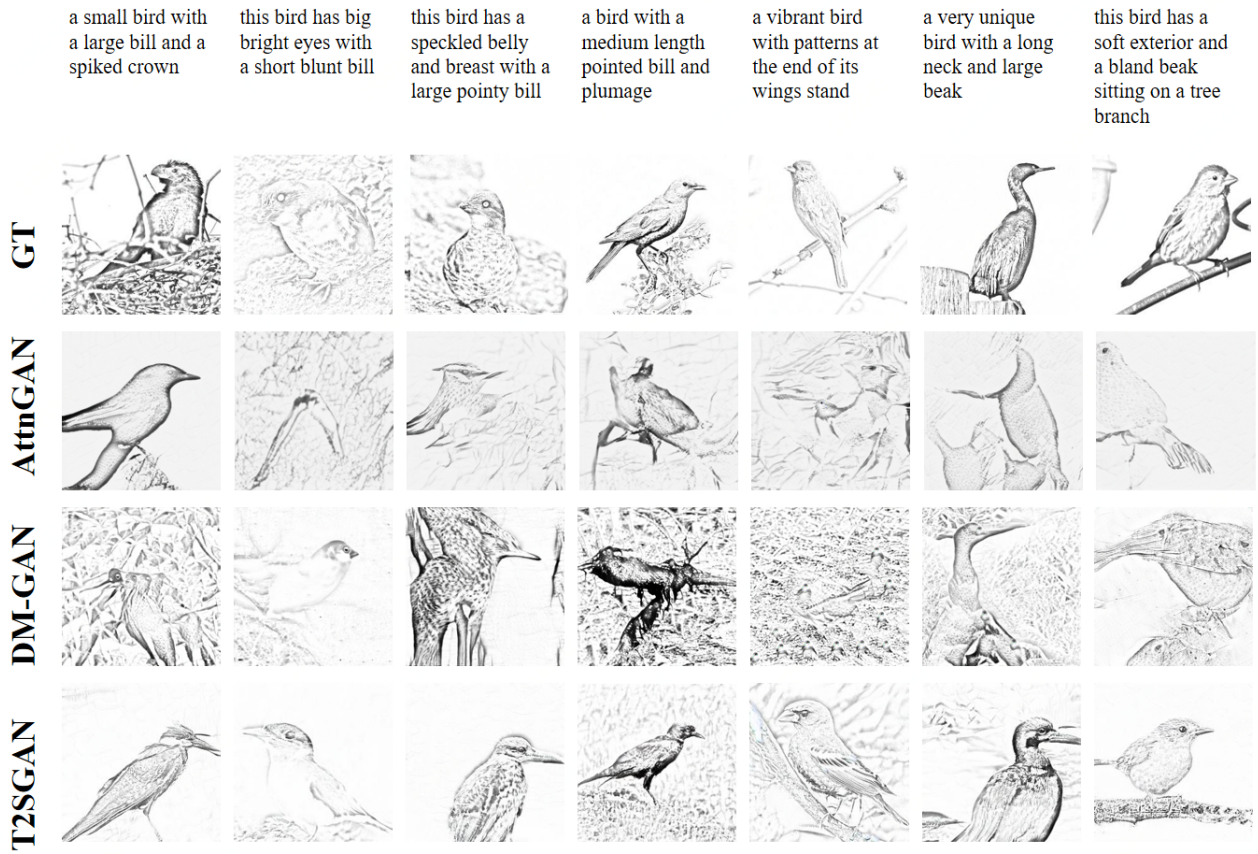

Figure 4. Generated sketch examples of different models.

It's observed that the sketch images generated by AttnGAN [31] and DM-GAN [37] look like simple shapes and some combinations of fuzzy lines. As shown in the 1st, 4th, 5th, 6th, and 7th columns, the sketches generated by AttnGAN [31] and DM-GAN [37] are difficult to depict the contours of birds. Benefited from the CLIN Fusion Blocks, our model can generate better object shapes and vivid feather details of birds (2nd, 5th, 6th, and 7th columns).

We also find the T2S-GAN can generate sketch images with better semantic consistency between text and sketch. For example, in the 1st, 2nd, 5th, and 7th columns of Figure 4, the baseline models can not present detailed information like "spiked crown", "blunt bill", "pattern at the end of its wings", and "bland beak", but our T2S-GAN can catch these patterns and keep consist with the text descriptions.

Table 2. Evaluation with different models on sketch generation.

Model	FID	SSIM
AttnGAN	55.55	0.41
DM-GAN	33.51	0.36
T2S-GAN(CBN)	31.42	0.41
T2S-GAN(CLIN)-$\mathcal{L}_{per}$	26.81	0.44
T2S-GAN(CLIN)-$\mathcal{L}_{dam}$	27.53	0.43
T2S-GAN(CLIN)	**26.20**	**0.45**

4.5. Human Evaluation

Table 3. The human evaluation results of different models.

Model	VQ Score	SC Score
AttnGAN	26.4	25.2
DM-GAN	31.2	29.5
T2S-GAN	**42.4**	**45.3**

To measure the quality of the generated images, we also conduct a user study to compare the user preference for our method and baseline approaches. We invite 15 graduate students majored in fine arts and with experience of sketch drawing. Given a pair of description text and reference sketch images, the evaluators are asked to pick one from three candidate images, while the model name is hidden. They are instructed to pick the most visually pleasing image and the one most consistent to the text descriptions semantically. Totally 200 generated sketch images were sam-

pled from each model for human evaluation. As shown in Table 3, our model obtains the majority of votes in both visual quality and content relevance. The results indicates that T2S-GAN can better capture the semantic information from the given descriptions and generate high-quality sketch images than the baselines.

4.6. Model Tuning

Table 4. Comparison of different numbers of CLIN-B blocks

Model	B2	B3	B4	B5	B6	B7	B8
FID	35.8	31.5	26.9	**26.2**	26.8	27.3	28.2

We conduct experiments to investigate how the number of CLIN-B blocks affect our model's performance. The experimental results are shown in Table 4, from which we can observe that, generally larger number of CLIN-B blocks result in better performances within certain range. From CLIN-B2 to CLIN-B5, the FID score decreases from 35.8 to 26.2. From CLIN-B5 to CLIN-B8, the FID score increases from 26.2 to 28.2. The quantitative comparisons of FID show that our model T2S-GAN reaches its best performance at CLIN-B5.

4.7. Error Analysis

Although T2SGAN can generate high-quality images that match the text, it also has some limitations. We demonstrate some typical bad cases of the generated sketches in Figure 5 to explore the limitations of our model.

1) *Under-interpreting:* As shown in Figure 5 (a), the bird stands on **a tree** as described in the given context. Whereas in the generated sketch, the bird is surrounded with blurring points. This is caused by the under-interpretation of the given text context, when there are multiple objects that need to be generated. In the future, we may introduce new loss function to encourage model to focus on multiple objects in the mentioned descriptions.

2) *Over-interpreting:* When there involves rich description for the features of the object, the model tends to focus mainly on certain feature and ignore the others. As shown in Figure 5 (b) (c), the context description includes detailed information of the birds such as *spiky plumage* or *speckled appearance*. However, the model tends to concentrate on features of *long tail* or *long neck* instead of generating the whole bird picture. One possible reason is that these short text description can not depict sketch completely. In future, we will enrich the description of sketch and improve the completeness of generated images.

3) *Mis-understanding:* Semantic interpretation plays an important role in sketch generation. Whereas in Figure 5 (d), the model fails to capture the semantic meaning of *glossy*. This mis-interpretation influences the stroke's clarity, and makes the object strokes blurry. To solve this problem, we will consider using more advanced text encoder such as Transformer [26] in the future.

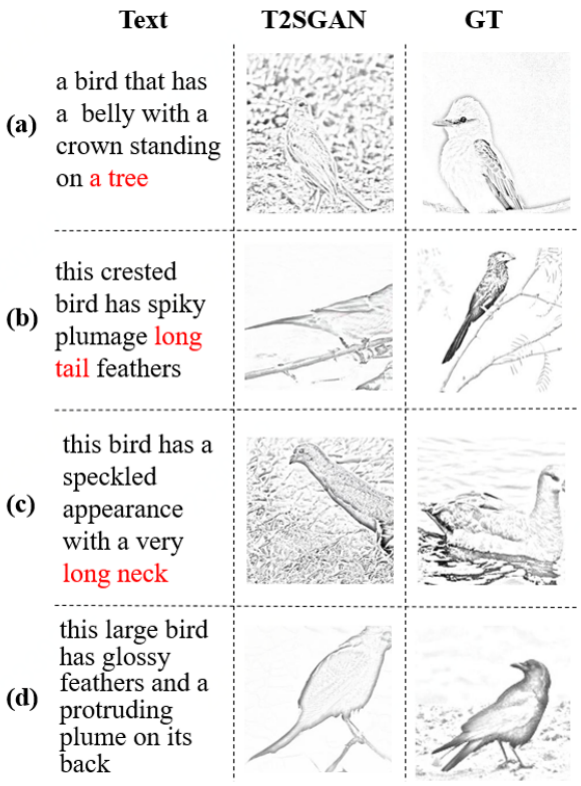

Figure 5. Some erroneous sketches generated by our model.

5. Conclusions

In this paper, we propose a novel multi-modal text-to-image task, which is to generate sketch images from natural language descriptions. Compared to existing text-to-image tasks, text-to-sketch requires the model to depict very accurate edge information and is more challenging. To facilitate the research, we construct an image dataset by modifying the classic CUB dataset. To fuse the muti-modal information of text and image effectively, we devise a novel GAN-based model T2SGAN which is equipped with several Conditional Layer-Instance Normalization (CLIN) based fusion blocks. The experimental results demonstrate the proposed model can catch the fine-grained patterns from text descriptions precisely and generate visually pleasing sketch images. We also did some error analysis to explore the future directions of this task.

References

[1] Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceed-*

ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9416–9425, 2018. 6

[2] Piotr Dollár and C Lawrence Zitnick. Fast edge detection using structured forests. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1558–1570, 2014. 6

[3] Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. Sketchycoco: Image generation from freehand scene sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5174–5183, 2020. 2

[4] Xinbo Gao, Nannan Wang, Dacheng Tao, and Xuelong Li. Face sketch–photo synthesis and retrieval using sparse representation. *IEEE Transactions on circuits and systems for video technology*, 22(8):1213–1226, 2012. 2

[5] Songwei Ge, Vedanuj Goswami, C Lawrence Zitnick, and Devi Parikh. Creative sketch generation. *arXiv preprint arXiv:2011.10039*, 2020. 3

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[7] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005. 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[9] 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. *Advances in neural information processing systems*, 30, 2017. 6

[10] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 6

[11] Kai Hu, Wentong Liao, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware gan. *arXiv preprint arXiv:2104.00567*, 2021. 2

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 6

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[14] Hyung W Kang, Wenjie He, Charles K Chui, and Uday K Chakraborty. Interactive sketch generation. *The Visual Computer*, 21(8):821–830, 2005. 1, 2

[15] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019. 2, 4

[16] Jia Li, Nan Gao, Tong Shen, Wei Zhang, Tao Mei, and Hui Ren. Sketchman: Learning to create professional sketches. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3237–3245, 2020. 2

[17] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2871, 2017. 3

[18] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016. 2, 4

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[20] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016. 6

[21] Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, and Zhijiang Zhang. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3982–3991, 2015. 6

[22] Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning to sketch with shortcut cycle consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 801–810, 2018. 1

[23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 6

[24] Hao Tang, Xinya Chen, Wei Wang, Dan Xu, Jason J Corso, Nicu Sebe, and Yan Yan. Attribute-guided sketch generation. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019. 2

[25] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020. 1, 2, 4

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 8

[27] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 5

[28] Yikun Wang, Liang Chang, Yuhua Cheng, Lihua Jin, Zhengxin Cheng, Xiaoming Deng, and Fuqing Duan. Text2sketch: Learning face sketch from facial attribute text. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 669–673. IEEE, 2018. 3

[29] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 2

[30] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 5

[31] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2, 3, 4, 5, 6, 7

[32] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2019. 2

[33] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2327–2336, 2019. 4, 5

[34] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2, 3

[35] Shengchuan Zhang, Xinbo Gao, Nannan Wang, Jie Li, and Mingjin Zhang. Face sketch synthesis via sparse representation-based greedy search. *IEEE transactions on image processing*, 24(8):2466–2477, 2015. 2

[36] Zhenxing Zhang and Lambert Schomaker. Dtgan: Dual attention generative adversarial networks for text-to-image generation. *arXiv preprint arXiv:2011.02709*, 2020. 5

[37] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019. 2, 5, 6, 7