Semi-supervised reference-based sketch extraction using a contrastive learning framework

CHANG WOOK SEO, Visual Media Lab, KAIST, South Korea
AMIRSAMAN ASHTARI, Visual Media Lab, KAIST, South Korea
JUNYONG NOH, Visual Media Lab, KAIST, South Korea

Fig. 1. Input color images and sketches extracted by our method. Without requiring repetitive training of the network to make pre-trained weights for each style, our model produces various style sketches by imitating the input reference sketches.© 4SKST (1,2,4), DICC (3), Comet_atr (5)

Sketches reflect the drawing style of individual artists; therefore, it is important to consider their unique styles when extracting sketches from color images for various applications. Unfortunately, most existing sketch extraction methods are designed to extract sketches of a single style. Although there have been some attempts to generate various style sketches, the methods generally suffer from two limitations: low quality results and difficulty in training the model due to the requirement of a paired dataset. In this paper, we propose a novel multi-modal sketch extraction method that can imitate the style of a given reference sketch with unpaired data training in a semi-supervised manner. Our method outperforms state-of-the-art sketch extraction methods and unpaired image translation methods in both quantitative and qualitative evaluations.

CCS Concepts: • **Applied computing** → *Fine arts*; • **Computing methodologies** → **Image processing**.

Authors' addresses: Chang Wook Seo, Visual Media Lab, KAIST, South Korea, lgtwins@kaist.ac.kr; Amirsaman Ashtari, Visual Media Lab, KAIST, South Korea, a.s.ashtari@kaist.ac.kr; Junyong Noh, Visual Media Lab, KAIST, South Korea, lgtwins@kaist.ac.kr.

0730-0301/2023/8-ART56 $15.00
https://doi.org/10.1145/3592392

Additional Key Words and Phrases: Sketch-extraction, Auto-colorization, Image-to-image translation

ACM Reference Format:
Chang Wook Seo, Amirsaman Ashtari, and Junyong Noh. 2023. Semi-supervised reference-based sketch extraction using a contrastive learning framework. *ACM Trans. Graph.* 42, 4, Article 56 (August 2023), 12 pages. https://doi.org/10.1145/3592392

1 INTRODUCTION

Sketches can be used for a variety of different purposes. Sometimes a sketch can be art by itself and other times it provides a glimpse of the final drawing as an intermediate step. They can also be used to deliver the thoughts of artists as an effective medium for visual communication. For example, artists draw a sketch with relatively complex and thick lines to convey strong impressions. If harmony and balance are intended between the sketch and colors, use of thin and abstract lines is often preferred.

Many computer vision and graphics studies have attempted to automatically extract sketches from photos [Ashtari et al. 2022; Chan et al. 2022] or generating in abstracted lines [Mo et al. 2021; Vinker et al. 2022]. During the extraction process, it is important to infuse the extracted sketches with the style of the authentic drawings, which would have been produced by artists for the outcome to be aesthetically pleasing. From this perspective, most widely used sketch extraction or edge-detection techniques such as Canny [1986]

and XDoG [Winnemöller 2011] do not serve the purpose as their results are often noisy or consist of dotted lines.

To address this, SketchKeras [lllyasviel 2017] utilizes deep learning to generate pencil stroke style sketches that closely imitate artistic sketches. Similarly, Chan et al. [2022] proposed a method that can produce high quality artistic sketch drawings by incorporating the geometry and semantics of a color image. Unfortunately, these approaches focus on generating a single style sketch. Ref2sketch [Ashtari et al. 2022] is a multi-modal method that can extract artistic sketches in various styles. Provided with a reference sketch as an additional input, the extracted Ref2sketch sketch closely reflects the style represented in the reference. However, this approach requires a large number of paired sketch and color image data because it is designed to learn the sketch style in a supervised manner.

In this paper, we propose a new sketch extraction method that learns to imitate the style of a reference sketch in the same way as Ref2sketch. However, by leveraging a pre-trained contrastive model, which was trained on paired data, our method can be trained using unpaired sketch and color image data. This approach enhances the efficiency of the training process in a semi-supervised manner. In addition, incorporating attention concatenation that emphasizes the spatial and channel information of inputs improves the quality of the produced sketch. To reflect the style of a reference sketch effectively, we propose a new sketch style loss that utilizes pre-trained weights. These weights are trained based on contrastive learning with a sketch dataset of various styles. We also adopt a line loss by utilizing the HED [Xie and Tu 2015] method to help the model generate a clear and accurate line shape.

Our contributions can be summarized as follows.

- We propose a novel multi-modal sketch extraction method that can imitate the drawing style of the input reference sketch. Our model is trainable with unpaired sketch and color images in a semi-supervised manner.
- We show how generated sketches can be utilized for related studies such as auto-colorization and sketch style transfer.
- In addition to the code, we provide a new authentic sketch dataset prepared by a professional artist. This dataset can assist in precisely evaluating various sketch extraction models. The dataset consists of one of four different styles of sketch drawings paired to 25 color images. The dataset includes a total number of 100 image pairs.

2 RELATED WORK

2.1 Sketch extraction

There are many sketch extraction techniques designed to generate corresponding sketch images from color images. Some approaches employ an edge-detection method such as Canny [1986], XDoG [Winnemöller 2011], or HED [Xie and Tu 2015]. Other approaches such as SketchKeras [lllyasviel 2017], Anime2sketch [Xiaoyu Xiang 2021], manga line extraction [Li et al. 2017a] and Sketch-simplifications [Simo-Serra et al. 2018, 2016; Xu et al. 2021] have the specific purpose of achieving high quality sketch images using deep learning. Recently, Chan et al. [2022] proposed a novel sketch extraction network that utilizes the depth and semantic meanings of the color image to visualize high quality sketch line drawings. Ref2sketch [Ashtari et al. 2022] is a multi-modal sketch extraction network that learns to imitate an input reference sketch to generate high quality artistic sketch outputs.

These learning-based methods utilize additional loss functions and layers on top of the network models introduced in general domain image-to-image translation studies to improve the performance specifically in the sketch domain. Our method utilizes an attention concatenation layer as well as a set of new loss functions to produce higher quality sketch images compared to previous studies. In addition, our model is trained with unpaired data in a semi-supervised manner to produce a sketch of the style given in the reference image.

2.2 Image-to-image translation

Image-to-image translation methods can be divided into several categories: supervised [Ashtari et al. 2022; Isola et al. 2017; Rott Shaham et al. 2021; Wang et al. 2018b,a], unsupervised [Kim et al. 2017; Nizan and Tal 2020; Park et al. 2020a; Xie et al. 2021; Zhu et al. 2017], single-modal [Isola et al. 2017; Rott Shaham et al. 2021; Wang et al. 2018b,a; Xie et al. 2021], and multi-modal [Choi et al. 2020; Lee et al. 2020b; Nizan and Tal 2020; Park et al. 2020b; Ruan et al. 2019]. Supervised methods require paired data for training the model, while unsupervised methods can be trained with unpaired data. Paired data is valuable but rare, especially for authentic sketches drawn by artists; therefore, unsupervised methods make provide convenience for dataset gathering more convenient.

Single-modal methods generate only one output for a given input, while multi-modal methods produce various outputs from either a single input or multiple additional inputs such as a segmentation map [Ntavelis et al. 2020; Sushko et al. 2020; Tang et al. 2022, 2019], text [Kim and Ye 2021; Li et al. 2020a,b,c; Liu et al. 2020a] or a reference image [He et al. 2018; Huang et al. 2018; Li et al. 2022; Ma et al. 2018a; Park et al. 2020b]. Multi-modal methods can easily be applied to diverse applications that require different style images in the same domain, including interior [Lee et al. 2017; Xu et al. 2017; Zheng et al. 2020], human-pose [Schneider et al. 2022; Si et al. 2018; Siarohin et al. 2018], and face emotions [Luna-Jiménez et al. 2021; Savchenko 2022; Seo et al. 2022]. Because different artists draw sketches in different styles, it is important to consider these style differences in sketch based image manipulation studies [Liu et al. 2021; Qi et al. 2021; Simo-Serra et al. 2016; Xu et al. 2021]. Therefore, multi-modal methods have been adopted in various sketch domain studies including photo-sketch synthesis [Chen and Hays 2018; Gao et al. 2020; Li et al. 2019, 2017b; Liu et al. 2020b; Yi et al. 2019, 2020] and sketch auto-colorization [Ci et al. 2018; Kim et al. 2019; Lee et al. 2020a; Liu et al. 2022; Ma et al. 2018b; Thasarathan and Ebrahimi 2019; Yuan and Simo-Serra 2021; Zhang et al. 2018b,b; Zou et al. 2019].

3 METHOD

Our goal is to design a model that extracts a sketch from a given color image while imitating the style of a reference sketch. Because pairs of sketch and color image data are scarce, we choose to train the model with an unpaired dataset. Many previous approaches

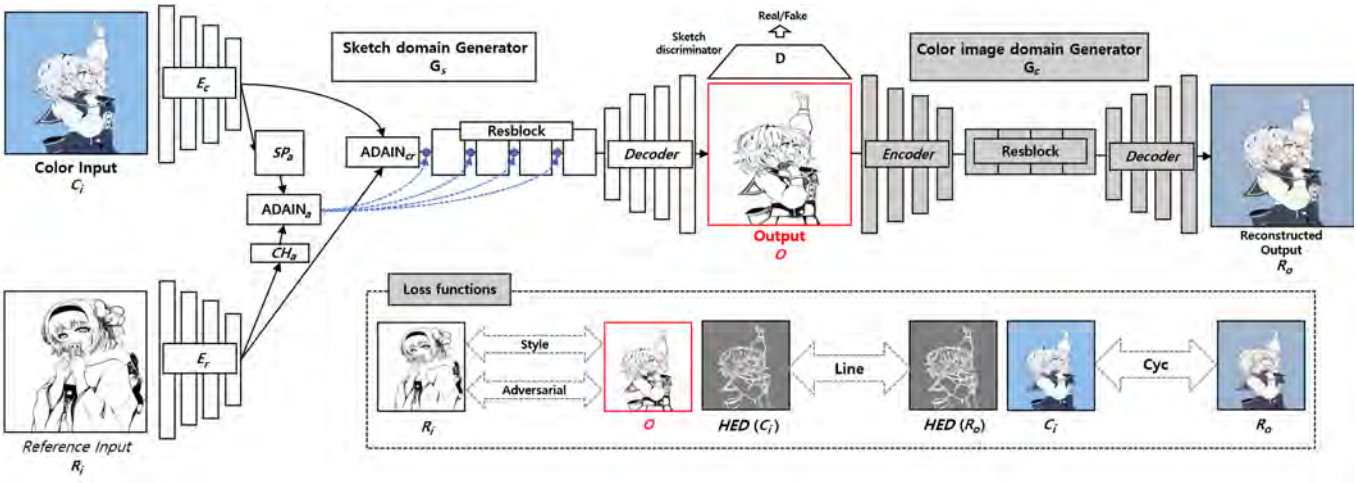

Fig. 2. Overview of our network design. See Section 3.1 for the explanation of this network and the definitions for the notations used here. © 4SKST

have relied solely on cycle consistency losses [Zhu et al. 2017] in the generator to preserve visual similarity between unpaired data. In addition to using the cycle consistency loss, we introduce two novel losses. The first is line loss that ensures the shape of the output sketch is similar. to that of the input color image. The second is a sketch style loss that enforces the style of the output sketch to follow that of the reference image. The discriminator D of our network examines if the output of the sketch domain generator $O{=}G_s(C_i, R_i)$ is in the same domain as that of the reference input R_i. Here, C_i represents the input color image. D ensures that O lies in a sketch domain. See Figure 2 for an overview of our network design and the supplementary material for the detailed information of our network.

3.1 Overview

Our method performs the following steps to train our sketch-extraction model that produces the output with the style defined by the reference sketch:

(A) Two encoders, E_c and E_r which consist of convolution layers, extract the features from the input color image C_i and input reference image R_i, respectively.
(B) The extracted features from E_c are fed into the spatial attention layer SP_a and features from E_r are fed into the channel attention layer CH_a. The outputs from the attention layers then go through adaptive instance normalization $ADAIN_a$ [Huang and Belongie 2017].
(C) Simultaneously with step (B), the features from E_c and E_r directly go through another adaptive instance normalization $ADAIN_{cr}$.
(D) Resblock, which consists of 4 convolution block layers, receives the outputs from both $ADAIN_a$ and $ADAIN_{cr}$. The output from $ADAIN_a$ is concatenated to the output from $ADAIN_{cr}$ before the combination is concatenated to the output of each Resblock layer.
(E) The output from Resblock goes through the decoder layers to produce an output sketch.
(F) To preserve the shape of the color image input in the output sketch, the output sketch is fed into the color image domain generator G_c. G_c, which consists of the encoder-decoder and Resblock layers, produces a reconstructed output R_o. We calculate four different loss functions using O and R_o to ensure that the output sketch imitates the style of the reference input while preserving the shape of the color image input.

3.2 Attention

Our network includes spatial attention and channel attention to emphasize the shape and style of each input. The attention structures are similar to those of the CBAM [Woo et al. 2018] method. To reflect the shape given by C_i, the spatial attention is placed after E_c. Likewise, to adopt the style given by R_i, the channel attention is placed after E_r. The details of the attentions are as follows.

For input feature maps, $E_c(C_i) \in \mathbb{R}^{C\times H\times W}$ and $E_r(R_i) \in \mathbb{R}^{C\times H\times W}$, where C, H, and W represent the number of channels, the height and the width of the image, respectively. We compute the spatial and channel attentions individually (i.e., $SP_a \in \mathbb{R}^{1\times H\times W}$ and $CH_a \in \mathbb{R}^{C\times 1\times 1}$) as follows:

$$\begin{aligned} &SP_a(E_c(C_i)) \\ &= \sigma(f^{3\times 3}([AvgPool^{sp}(E_c(C_i)); MaxPool^{sp}(E_c(C_i))])) \\ &= \sigma(f^{3\times 3}([E_c(C_i)^{sp}_{avg}; E_c(C_i)^{sp}_{max}])), \end{aligned} \tag{1}$$

$$\begin{aligned} &CH_a(E_r(R_i)) \\ &= \sigma(MLP(AvgPool^{ch}(E_r(R_i))) + MLP(MaxPool^{ch}(E_r(R_i)))) \\ &= \sigma(W_1(W_0(E_r(R_i)^{ch}_{avg})) + W_1(W_0(E_r(R_i)^{ch}_{max}))), \end{aligned} \tag{2}$$

In Eq. (1), the features from E_c are pooled by two different pooling functions before convolved with a 3×3 kernel filter. In Eq. (2), the features from E_r are pooled by two different pooling functions before

going through MLP layers. Here, $W_0 \in \mathbb{R}^{C/r \times 1}$ and $W_1 \in \mathbb{R}^{C \times C/r}$ represent the weights of MLP layers, with a reduction ratio $r = 16$. Sigmoid functions σ are used for both attentions. The feature sizes after the pooling layers are $AvgPool^{ch}, MaxPool^{ch} \in \mathbb{R}^{C \times 1 \times 1}$ and $AvgPool^{sp}, MaxPool^{sp} \in \mathbb{R}^{1 \times H \times W}$.

These attention features are then multiplied element-wise using the Hadamard product $\odot$ with their original input features $E_c(C_i)$ and $E_r(R_i)$ before being normalized by $ADAIN_a$. ADAIN aligns the mean and variance of the features from the color and reference inputs:

$$ADAIN_a(E_c(C_i) \odot SP_a(E_c(C_i)), E_r(R_i) \odot CH_a(E_r(R_i))). \quad (3)$$

These normalized features are concatenated to the features from $ADAIN_{cr}$ at Resblock to make the model trainable at a higher resolution. Training the model at a low resolution such as 256×256 without the attention feature concatenation preserves the shape but produces low-quality sketches. Training the model at a higher resolution such as 512×512 without the concatenation causes shape distortion. See Figure 3 for examples. More discussion on the benefit of this attention design can be found in the supplementary material.

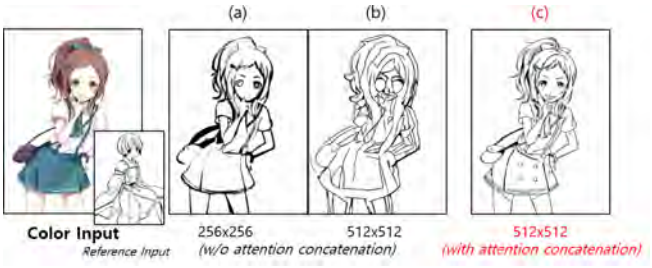

Fig. 3. Examples of the output produced with and without attention concatenation in the network architecture. For this experiment, we removed SP_a, CH_a, and $ADAIN_a$ as well as their connections to Resblock layers. The outputs produced without the attention concatenation, (a) and (b), show poor quality results while the output using the concatenation shows a better result produced at a higher resolution of 512×512. © 4SKST

3.3 Losses

3.3.1 Sketch Style Loss. Here, we introduce a novel loss function that calculates the style difference in the sketch domain. This loss is computed using weights pre-trained based on contrastive learning [Chen et al. 2020] that employs a triplet loss. Contrastive learning maps similar features closer together and dissimilar features further away in embedded space.

In our method, the model is trained to map sketches of similar styles closer together (Anchor and Positive) and sketches of the same shape but different styles (Negative) further away (as illustrated in Figure 4). The dataset for training the network is obtained using Ref2sketch [Ashtari et al. 2022], which allows sketches of different styles to be generated from a single input image.

To compute the style loss, the reference input R_i and output $O = G_s(C_i, R_i)$ generated by the sketch domain generator G_s are used. Although these two images are of different shapes, they should be of the same style; therefore, we extract the style feature embeddings from them using pre-trained contrastive weights. We then apply L1 normalization to calculate the difference of the two embeddings. The loss is expressed as follows:

$$\mathcal{L}_{style} = \mathbb{E}_{O,R_i}[||C_w(O) - C_w(R_i)||_1]. \quad (4)$$

Symbol C_w represents the pre-trained contrastive learning model that extracts style feature embeddings. Without this style loss function, the network fails to imitate the drawing style of the reference input and consequently produces a fixed style output regardless of the reference input. Figure 5 shows failed examples. Similar to Simo-Serra et al. [2018], this loss function is pre-trained with paired data to enable our main model to be trained with unpaired data. This makes our approach semi-supervised. Refer to the supplementary material for more details regarding the contrastive learning weights used in our method.

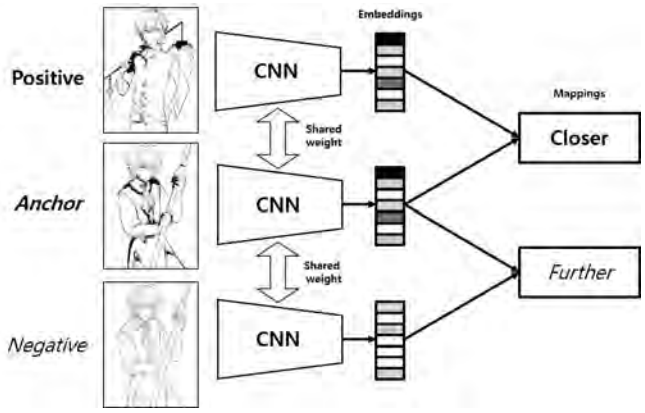

Fig. 4. Illustration of contrastive learning, which groups similar sketch styles in embedded space regardless of the image shape. © 4SKST

3.3.2 Line Loss. To enforce the shape of the output to be identical to the color image input, we apply a loss function that compares the edges of the reconstructed output $R_o = G_c(O)$ generated by the color domain generator G_c and the color input C_i using the HED [Xie and Tu 2015] method. HED [Xie and Tu 2015] detects the edges from the input image. The loss is expressed as follows:

$$\mathcal{L}_{Line} = \mathbb{E}_{C_i,R_o}[\sum_l ||\phi_l(HED(C_i)) - \phi_l(HED(R_o))||_1], \quad (5)$$

Applying HED to the color input C_i and reconstructed output R_o generates edge-detected images, as shown in Figure 2. The differences between the two edge-detected images, $HED(C_i)$ and $HED(R_o)$, are calculated by the perceptual loss function [Johnson et al. 2016] that is designed to compare images based on the pre-trained VGG16 [Simonyan and Zisserman 2014] model. ϕ_l denotes the activation map from the l^{th} layer of the VGG16 network. Without this line loss function, the network produces a shape that looks different from that of the color image input. Figure 5 shows failed examples, particularly on the dog's muzzle. This loss function is inspired by Yi et al. [2020].

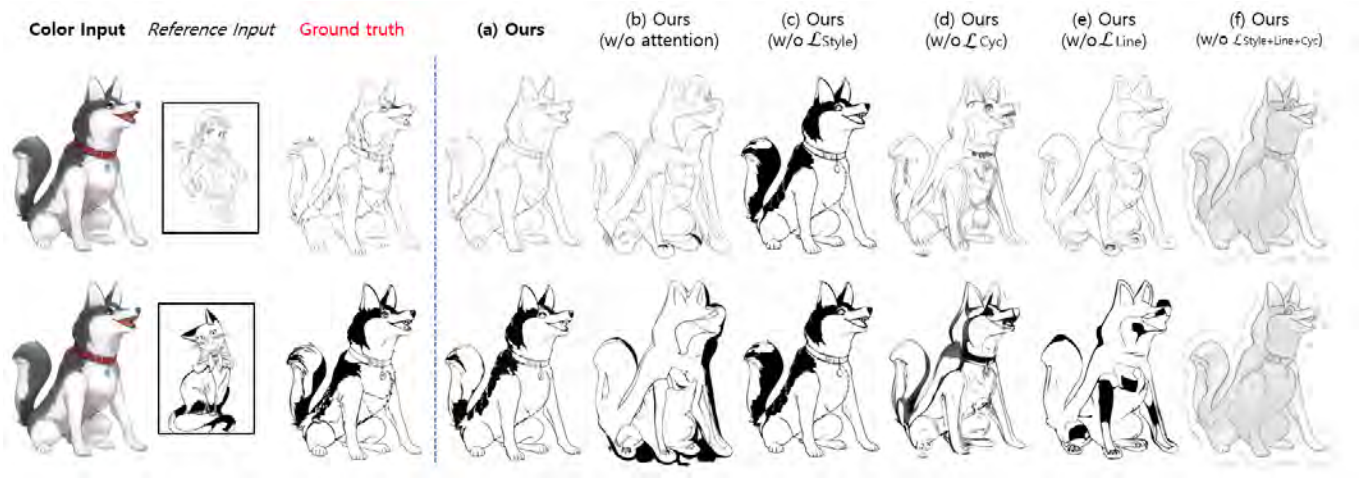

Fig. 5. Examples of outputs produced with and without using loss functions. (c) and (f) show the results from the model trained without using the style loss. It is evident that the results have a fixed style regardless of the given reference input. (a), (b), (d), and (e) show the results produced using the style loss. Clearly, the results displayed in the top and bottom rows have different styles according to the reference input. (e) shows the results from the model trained without using the line loss. The shape of the color image input is incorrectly represented. (d) shows the results trained with the line loss but without the cycle consistency loss. The image shapes of (d) are better than (e) but still do not correctly represent the original shape. © 4SKST

3.3.3 Cycle Consistency Loss. We enforce the shape similarity further by comparing the overall visual difference between C_i and R_o. L1 normalization is used, and the loss is expressed as follows:

$$\mathcal{L}_{Cyc} = \mathbb{E}_{C_i,R_o}[||C_i - R_o||_1], \tag{6}$$

3.3.4 Adversarial Loss. The adversarial loss forces the discriminator D to classify the synthetic output O to be in the domain similar to that defined by R_i, which is a sketch in our case. The loss is expressed as follows:

$$\mathcal{L}_{adv} = \mathbb{E}_{R_i}[\log(D(R_i))] + \mathbb{E}_O[\log(1 - D(O))], \tag{7}$$

The total loss function for the generator G and discriminator D is defined as follows:

$$\min_G \max_D \mathcal{L}_{total} = \lambda_{style}\mathcal{L}_{style} + \lambda_{line}\mathcal{L}_{line} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{adv}\mathcal{L}_{adv}. \tag{8}$$

The parameters used in Eq. 8 are $\lambda_{style,line}$ = 5 - $\frac{4.5i}{n}$, λ_{cyc} = 10, and λ_{adv} = 1, where i is the current epoch number and n is the total number of epochs.

4 EXPERIMENT

4.1 Experiment setup

4.1.1 Dataset. To train our network and baselines, we utilized images from safebooru [DanbooruCommunity 2021]. For sketch domain images, we used images with the tag *line art* in safebooru.

A total of 4,302 sketch domain images are used for training. For the color domain, 3,804 color images were randomly selected from safebooru, which are not used as sketch data. Refer to the supplementary material for detailed information regarding the training dataset.

To evaluate our method and baselines, we created a new sketch dataset of four different styles with the help of a professional artist. Recently, Ref2sketch [Ashtari et al. 2022] evaluated their model and baselines with sketches of four different styles, which consists of a total of 60 pairs. Similarly, we performed the evaluation with a dataset of four different styles to verify how well the output of each model imitates the given reference style. Our **4 sketch style (4SKST) dataset** consists of each of four different style sketches for 25 color images, constituting a total of 100 pairs. Color images were chosen from two different domains: anime-art and real photo. These images and sketches are free to re-distribute for non-commercial purposes (CC-NC)[1]. Four different sketch drawing styles in the dataset follow four major sketch drawing styles. The four major styles were determined by applying K-means clustering [Lloyd 1982] to the same 4,302 sketch images used for training.

4.1.2 Training details. An Adam optimizer [Kingma and Ba 2014] with a batch size of four was used in the training. All networks were trained from scratch with a learning rate of 0.0002, and the total number of training epochs was 100 with a constant learning rate for the first 50 epochs followed by a learning rate linearly decayed to zero over the next 50 epochs.

4.2 Ablation study

For an ablation study, our network was trained with the attention layers and losses removed according to Figure 5. The evaluation of each model was performed using the 4SKST dataset. When extracting a sketch from a color image input, the same sketch style was used; however, unseen shape images from the dataset were used as the input reference sketch. As evaluation metrics, we employed PSNR [Wang et al. 2004], FID [Heusel et al. 2017], and LPIPS [Zhang

[1] CreativeCommons for Non-Commercial uses

et al. 2018a] to compare the distribution of features from the output and ground truth sketches. Table 1 shows the results and Figure 5 illustrates the examples.

The model trained without using attention concatenation performs very poorly and produces the worst results. The models trained without using the style loss, cycle consistency loss, or line loss also produce much different results than the ground truth. The results produced without simultaneously using these three losses have a visual quality far from the sketch domain and achieve the worst scores in both PSNR and FID.

4.3 Comparison with baselines

We chose six different baselines to compare with our method, MUNIT [Huang et al. 2018], Park et al. [2020b], Ref2sketch [Ashtari et al. 2022], CouncilGAN [Nizan and Tal 2020], IrwGAN [Xie et al. 2021], and Chan et al. [2022]. MUNIT and Park et al. [2020b] are unsupervised multi-modal image translation methods that imitate the reference input for the output style. Note that these methods are designed for general image domain translations. Ref2sketch is the most recent multi-modal method with supervised learning specifically designed for the sketch domain. Council-GAN and IrwGAN are unsupervised image translation methods that solve the limitation of cycle consistency learning based methods [Zhu et al. 2017] by leveraging the collaboration between GANs and the importance reweighting technique. Unfortunately, these methods do not accept a reference input to imitate, and thereby cannot produce a sketch of a desired style. Similarly, while Chan et al. [2022] can convey the semantic and depth meaning of the color image input to an output sketch in an unsupervised manner, it is a single-modal method that can produce the sketch of only one style unless the model is retrained with a dataset of different style sketches.

For comparison, we trained the baseline models with the same dataset described in Sec. 4.1. The training parameters and details were determined based on their official code and the descriptions in their respective papers. Because Ref2sketch [Ashtari et al. 2022] is a supervised method that requires a paired dataset, we utilized pre-trained weights from the official page [ref2sketch 2022]. In the evaluation of each model, methods that can accept a reference image [Ashtari et al. 2022; Huang et al. 2018; Park et al. 2020b] use the same style sketch with an unseen shape image from the 4SKST dataset. Other methods that cannot accept a reference image [Chan et al. 2022; Nizan and Tal 2020; Xie et al. 2021] generate the output based on the color image input. Figure 6 shows examples of the generated outputs. The output sketches produced by these methods were compared to the ground truth with the same three different evaluation metrics [Heusel et al. 2017; Wang et al. 2004; Zhang et al. 2018a] used for the ablation study. A total of 100 pairs of images from the 4SKST dataset were used for each evaluation. Refer to the supplementary material for a detailed explanation regarding the experiment. The results reported in Table 1 confirm that our method outperforms all the baselines in the three different evaluation metrics.

Table 1. Quantitative results from the ablation study and from the comparison with baselines.

Methods	PSNR↑	FID↓	LPIPS↓
Ours	**35.58**	**82.18**	**0.1271**
Ours w/o attention	33.80	146.87	0.3356
Ours w/o $\mathcal{L}_{Style}$	34.96	139.28	0.1738
Ours w/o $\mathcal{L}_{Cyc}$	34.68	121.71	0.2357
Ours w/o $\mathcal{L}_{Line}$	34.24	125.15	0.2660
Ours w/o $\mathcal{L}_{Style}$ + $\mathcal{L}_{Line}$ + $\mathcal{L}_{Cyc}$	33.59	157.92	0.2598
MUNIT	34.23	144.82	0.2582
Park et al. [2020b]	35.04	174.12	0.2745
Ref2sketch	35.02	115.96	0.2192
Council-GAN	31.76	215.81	0.4632
IrwGAN	35.36	125.14	0.2229
Chan et al. [2022]	35.05	128.96	0.2130

4.4 Perceptual study

We further evaluated the performance of our method based on human perceptual judgment. A total of 200 people participated in this study, and a survey consisting of 20 comparisons was created for our evaluation. We provided each participant with a target image and seven results, including ours, in each comparison. We then asked each participant to select the result that looks most similar to the target image. No time constraint was imposed on the participants in this process. Figure 7 shows an example survey and Table 2 lists the resulting scores. More examples can be found in the supplementary material. The result of this perceptual study clearly verifies that our method outperforms the baselines in human perception.

Table 2. Results from the user perceptual study

Method	User Score
Ours	**79.50%**
Ref2sketch	10.66%
Chan et al. [2022]	4.58%
IrwGAN	3.08%
MUNIT	1.16%
Park et al. [2020b]	1.02%
Council-GAN	0%

4.5 Cyclic evaluation

To prove the superiority of our method in preserving the style of the given reference when extracting sketches compared to other methods, we implemented the cyclic evaluation proposed in Ref2sketch [Ashtari et al. 2022]. The main idea of the cyclic evaluation is that, because the extracted sketch should have a style similar to that of the reference input, using the output as the reference image in turn will lead to the same sketch in the original style. Figure 8 illustrates this process. In this evaluation, we chose MUNIT [Huang et al. 2018], Park et al. [2020b], and Ref2sketch as baselines because these methods are designed to accept a reference input to imitate the style. The 4SKST dataset was used for the evaluation. As shown

Fig. 6. Various examples generated by our method and baselines. While Ref2sketch produces high quality results in some cases as shown in the 6th row, overall our method produces the best quality results in most cases. © 4SKST

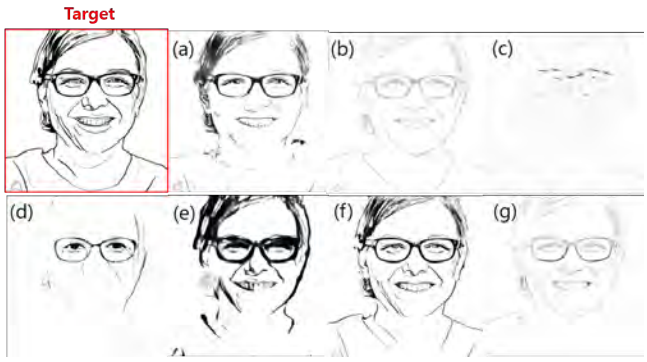

Fig. 7. Sample survey prompt for our human perception study. The presented results were produced by our method and six baseline methods: (a) Ref2sketch, (b) Chan et al. [2022], (c) Council-GAN, (d) Park et al. [2020b], (e) MUNIT, (f) Ours, and (g) IrwGAN. The displayed order of these results was chosen randomly for each comparison. © 4SKST

in Table 3, our method outperforms the baselines in LPIPS, FID, and PSNR scores.

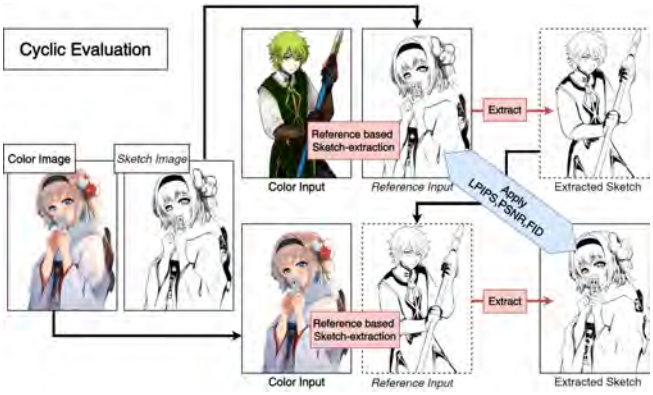

Fig. 8. Illustration of the cyclic evaluation proposed in Ref2sketch [Ashtari et al. 2022]. © 4SKST

Table 3. Results of the cyclic evaluation

Cylclic evaluation	PSNR↑	FID↓	LPIPS↓
Ours	**35.84**	**91.97**	**0.1288**
Ref2sketch	35.34	130.02	0.1643
Park et al. [2020b]	34.48	155.44	0.2824
MUNIT	33.92	151.04	0.2987

4.6 Applications

4.6.1 Improving auto-colorization. As described in Sec. 2.2, a sketch extraction method that can produce sketches of various styles can also be used to improve the performance of related applications. For example, sketch auto-colorization models typically require many sketches paired with color images [Kim et al. 2019; Lee et al. 2020a; Liu et al. 2022; Ma et al. 2018b; Thasarathan and Ebrahimi 2019; Yuan and Simo-Serra 2021; Zhang et al. 2018b]. These models can benefit from our method in that the use of a multi-style sketch dataset can help avoid an over-fitting problem caused by relying on the sketches of a single style for training.

Table 4. Comparison of the results from the auto-colorization method that was trained using different datasets. The "combined" category consists of sketches extracted by Canny, SketchKeras, XDoG and Simo-Serra et al. [2016].

Dataset	PSNR↑	FID↓	LPIPS↓
Ours	**29.40**	**227.69**	**0.4802**
Canny	28.68	268.10	0.5892
SketchKeras	27.99	283.69	0.6775
XDoG	28.55	239.63	0.5931
Simo-Serra et al. [2016]	28.49	270.99	0.6295
Combined	28.78	232.37	0.5604

To prove this, we trained an auto-colorization network with a paired sketch dataset which was generated by our method and baseline methods. A widely used deep-learning based sketch auto-colorization method [Ci et al. 2018] was chosen, and the model was trained with 1,500 color images from safebooru [DanbooruCommunity 2021], sketches generated by our network, and by the baseline methods used for the experiments performed in auto-colorization papers [Ci et al. 2018; Kim et al. 2019; Thasarathan and Ebrahimi 2019; Yuan and Simo-Serra 2021]. After training the model, we evaluated the quality of the auto-colorized images through authentic sketch inputs and the ground truth color images from the 4SKST dataset. Refer to the supplementary material for more detailed explanations regarding this experiment. Table 4 shows the comparison results.

The colorization model trained with the dataset generated by our method outperformed the others because the sketch dataset with various styles generated by our method helps avoid over-fitting to a specific single sketch style. We acknowledge that a similar effect would be achieved by training the model using the combined data from various sketch extraction methods, as Yuan and Simo-Serra [2021] attempted. These results are represented by "Combined" in Table 4. Our method still produces higher quality colorization results. It is also much simpler to extract multi-style sketches using our method than collecting data from many different methods. See Figure 9 for the colorized image examples.

4.6.2 Sketch style transfer. Similar to existing sketch style transfer methods [Liu et al. 2021; Simo-Serra et al. 2018, 2016; Xu et al. 2021], our method can transfer the style of a sketch directly to other sketches without extracting them from color images. This can be very useful when sketch artists work together. In the creation of comics, for example, character sketches are placed on template background sketches. Due to the involvement of many different artists, sketches are often prepared with different styles; therefore, this process requires a manually intensive arrangement of the same style sketches. Automatically transferring the style of a character sketch to match that of the background template will streamline

Fig. 9. Examples of the output from the auto-colorization method [Ci et al. 2018] trained with different datasets. © 4SKST

this collaboration process. Figure 10 shows an example of this application. This sketch style transfer application can also be used for

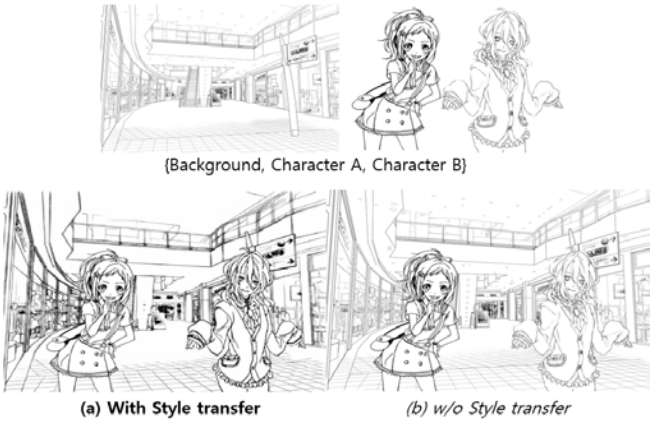

Fig. 10. Sketches of different styles can be placed easily in one scene after the style transfer by our method. Visually consistent (a) and inconsistent (b) placements of characters on the template background. © 4SKST (character), tzcat (background)

the purpose of sketch simplification [Simo-Serra et al. 2018, 2016; Xu et al. 2021]. Digitized rough sketches typically go through a simplification process for the purpose of cleaning up the image. Because it can accept a guiding reference for the simplification, our model can be more instrumental in creating cleaned up sketches in a desired uniform style compared to alternative methods. Refer to Figure 11 and the supplementary material for examples.

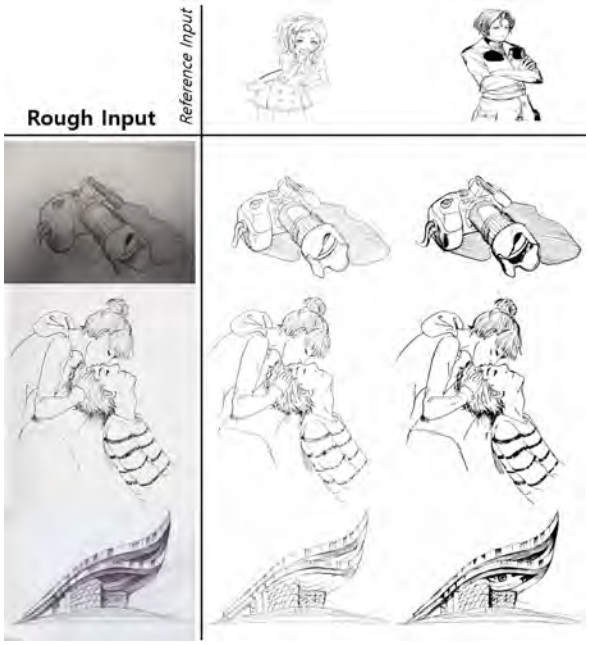

Fig. 11. Rough sketches can be simplified by transferring the style with our method. Example rough sketches are from Yan et al. [2020].

5 LIMITATIONS AND FUTURE WORK

Our method produces high-quality results when the categories of the color image and the reference image match. For example, higher quality landscape sketches will be extracted from a landscape photo

when a reference sketch with landscape content is provided. See Figure 12 for the examples. This unaligned image problem may be alleviated by adopting an importance reweighting method such as that proposed in IrwGAN [Xie et al. 2021]. How to apply this general image domain approach to the sketch domain is not yet clear and may be considered an interesting future research direction.

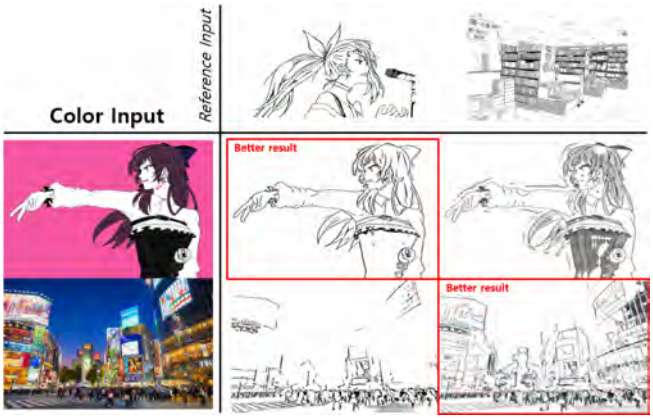

Fig. 12. When the categories of the reference and color images match, a higher quality sketch is extracted from the input color image. © 4SKST (character), tzcat (background)

Our method utilizes sketches extracted by Ref2sketch [Ashtari et al. 2022] when pre-training the contrastive learning model that works as the style loss function of our training network. Therefore, our method shares the same limitation with Ref2sketch. Specifically, our method cannot imitate the style of a reference sketch that does not consist of lines (e.g., pointillism art). The examples in Figure 13 illustrate this. This problem can be addressed by providing a dataset of these styles when pre-training the contrastive learning model.

6 CONCLUSION

In this study, we proposed a novel multi-modal method that can extract sketches from a color image in a style given by a reference image. Our method is trained efficiently in a semi-supervised manner. To imitate the style of the reference sketch, we used a pre-trained sketch style loss based on contrastive learning. The pre-training was performed with a paired dataset generated by Ref2sketch [Ashtari et al. 2022], which, in turn, was trained with another paired dataset. Leveraging the previous methods trained with paired data, our method was trained with unpaired color and sketch images.

To preserve the shape of the color input, we introduce a line loss function that is used on top of a cycle consistency loss function. Incorporation of the attention concatenation that emphasizes spatial and channel information enables our model to be trainable in high resolution, producing better quality results than those produced by the baselines. We verified the effectiveness of our method through both quantitative and qualitative experiments. We believe that our method can be utilized in the industry in the form of diverse applications and can stimulate further research related to the generation of various style sketches.

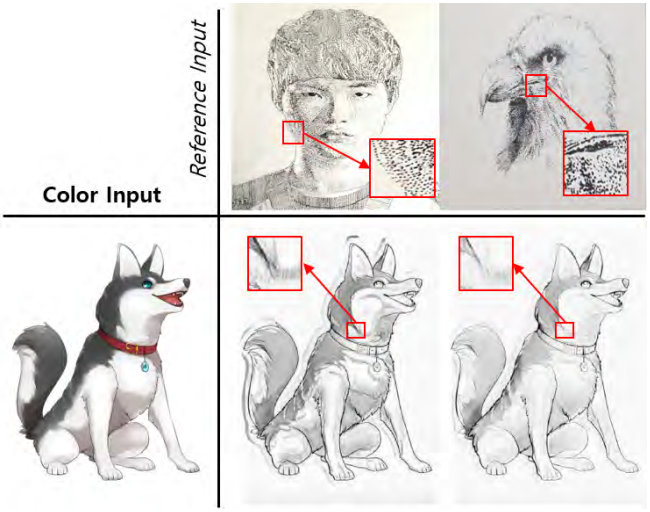

Fig. 13. Similar to Ref2sketch, our method fails to extract sketches of the reference style that does not consist of lines, such as pointillism art. © 4SKST (character), Kang (pointillism art)

ACKNOWLEDGMENTS

This work was supported by Korea Creative Content Agency (KOCCA) grant funded by the Korea government (MSIT) (Project Name: Development of universal fashion creation platform technology for avatar personality expression, Project Number: RS-2023-0022833). Special thanks to Minju Kim and Jungsuk Hur.

REFERENCES

Amirsaman Ashtari, Chang Wook Seo, Cholmin Kang, Sihun Cha, and Junyong Noh. 2022. Reference Based Sketch Extraction via Attention Mechanism. *ACM Trans. Graph.* 41, 6, Article 207 (nov 2022), 16 pages. https://doi.org/10.1145/3550454.3555504

J Canny. 1986. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 8, 6 (June 1986), 679–698. https://doi.org/10.1109/TPAMI.1986.4767851

Caroline Chan, Frédo Durand, and Phillip Isola. 2022. Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7915–7925.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

Wengling Chen and James Hays. 2018. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9416–9425.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8188–8197.

Yuanzheng Ci, Xinzhu Ma, Zhihui Wang, Haojie Li, and Zhongxuan Luo. 2018. User-guided deep anime line art colorization with conditional adversarial networks. In *Proceedings of the 26th ACM international conference on Multimedia*. 1536–1544.

DanbooruCommunity. 2021. Danbooru2020: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset. https://www.gwern.net/Danbooru2020. https://www.gwern.net/Danbooru2020 Accessed: 2022/04/03.

Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. 2020. Sketchycoco: Image generation from freehand scene sketches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5174–5183.

Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. 2018. Deep exemplar-based colorization. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–16.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran

Associates Inc., Red Hook, NY, USA, 6629–6640.
Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*. 1501–1510.
Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*. 172–189.
P. Isola, J. Zhu, T. Zhou, and A. A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5967–5976.
Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.
Gwanghyun Kim and Jong Chul Ye. 2021. Diffusionclip: Text-guided image manipulation using diffusion models. (2021).
Hyunsu Kim, Ho Young Jhoo, Eunhyeok Park, and Sungjoo Yoo. 2019. Tag2pix: Line art colorization using text tag with secat and changing loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9056–9065.
Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*. PMLR, 1857–1865.
Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (12 2014).
Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, and Andrew Rabinovich. 2017. Roomnet: End-to-end room layout estimation. In *Proceedings of the IEEE international conference on computer vision*. 4865–4874.
Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. 2020b. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision* 128, 10 (2020), 2402–2417.
Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. 2020a. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5801–5810.
Boyi Li, Serge Belongie, Ser-nam Lim, and Abe Davis. 2022. Neural Image Recolorization for Creative Domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2226–2230.
Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. 2020a. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7880–7889.
Bowen Li, Xiaojuan Qi, Philip Torr, and Thomas Lukasiewicz. 2020b. Lightweight generative adversarial networks for text-guided image manipulation. *Advances in Neural Information Processing Systems* 33 (2020), 22020–22031.
Chengze Li, Xueting Liu, and Tien-Tsin Wong. 2017a. Deep extraction of manga structural lines. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.
Jiguo Li, Xinfeng Zhang, Chuanmin Jia, Jizheng Xu, Li Zhang, Yue Wang, Siwei Ma, and Wen Gao. 2020c. Direct speech-to-image translation. *IEEE Journal of Selected Topics in Signal Processing* 14, 3 (2020), 517–529.
Yuhang Li, Xuejin Chen, Feng Wu, and Zheng-Jun Zha. 2019. Linestofacephoto: Face photo generation from lines with conditional self-attention generative adversarial networks. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2323–2331.
Yi Li, Yi-Zhe Song, Timothy M Hospedales, and Shaogang Gong. 2017b. Free-hand sketch synthesis with deformable stroke models. *International Journal of Computer Vision* 122, 1 (2017), 169–190.
Bingchen Liu, Kunpeng Song, Yizhe Zhu, and Ahmed Elgammal. 2020b. Sketch-to-art: Synthesizing stylized art images from sketches. In *Proceedings of the Asian Conference on Computer Vision*.
Xueting Liu, Wenliang Wu, Chengze Li, Yifan Li, and Huisi Wu. 2022. Reference-guided structure-aware deep sketch colorization for cartoons. *Computational Visual Media* 8, 1 (2022), 135–148.
Xueting Liu, Wenliang Wu, Huisi Wu, and Zhenkun Wen. 2021. Deep Style Transfer for Line Drawings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 353–361.
Yahui Liu, Marco De Nadai, Deng Cai, Huayang Li, Xavier Alameda-Pineda, Nicu Sebe, and Bruno Lepri. 2020a. Describe what to change: A text-guided unsupervised image-to-image translation approach. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1357–1365.
lllyasviel. 2017. sketchKeras. https://github.com/lllyasviel/sketchKeras.
Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE transactions on information theory* 28, 2 (1982), 129–137.
Cristina Luna-Jiménez, Ricardo Kleinlein, David Griol, Zoraida Callejas, Juan M Montero, and Fernando Fernández-Martínez. 2021. A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset. *Applied Sciences* 12, 1 (2021), 327.
Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. 2018a. Exemplar guided unsupervised image-to-image translation with semantic consistency. *arXiv preprint arXiv:1805.11145* (2018).
Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. 2018b. Exemplar guided unsupervised image-to-image translation with semantic consistency. *arXiv preprint arXiv:1805.11145* (2018).
Haoran Mo, Edgar Simo-Serra, Chengying Gao, Changqing Zou, and Ruomei Wang. 2021. General virtual sketching framework for vector line art. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–14.
Ori Nizan and Ayellet Tal. 2020. Breaking the cycle-colleagues are all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7860–7869.
Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. 2020. Sesame: Semantic editing of scenes by adding, manipulating or erasing objects. In *European Conference on Computer Vision*. Springer, 394–411.
Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. 2020a. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*. Springer, 319–345.
Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. 2020b. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems* 33 (2020), 7198–7211.
Yonggang Qi, Guoyao Su, Pinaki Nath Chowdhury, Mingkang Li, and Yi-Zhe Song. 2021. Sketchlattice: Latticed representation for sketch manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 953–961.
ref2sketch. 2022. Ref2sketch official page. https://github.com/ref2sketch/ref2sketch.
Tamar Rott Shaham, Michael Gharbi, Richard Zhang, Eli Shechtman, and Tomer Michaeli. 2021. Spatially-Adaptive Pixelwise Networks for Fast Image Translation. In *Computer Vision and Pattern Recognition (CVPR)*.
Congcong Ruan, Dihu Chen, and Haifeng Hu. 2019. Multimodal supervised image translation. *Electronics Letters* 55, 4 (2019), 190–192.
Andrey V Savchenko. 2022. HSEmotion: High-speed emotion recognition library. *Software Impacts* (2022), 100433.
David Schneider, Saquib Sarfraz, Alina Roitberg, and Rainer Stiefelhagen. 2022. Pose-based contrastive learning for domain agnostic activity representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3433–3443.
Kwanggyoon Seo, Seoung Wug Oh, Jingwan Lu, Joon-Young Lee, Seonghyeon Kim, and Junyong Noh. 2022. StylePortraitVideo: Editing Portrait Videos with Expression Optimization. (2022).
Chenyang Si, Wei Wang, Liang Wang, and Tieniu Tan. 2018. Multistage adversarial losses for pose-based human image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 118–126.
Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. 2018. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3408–3416.
Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. 2018. Mastering Sketching: Adversarial Augmentation for Structured Prediction. *ACM Trans. Graph.* 37, 1, Article 11 (jan 2018), 13 pages. https://doi.org/10.1145/3132703
Edgar Simo-Serra, Satoshi Iizuka, Kazuma Sasaki, and Hiroshi Ishikawa. 2016. Learning to simplify: fully convolutional networks for rough sketch cleanup. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–11.
Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. 2020. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781* (2020).
Hao Tang, Philip HS Torr, and Nicu Sebe. 2022. Multi-Channel Attention Selection GANs for Guided Image-to-Image Translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2022).
Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J. Corso, and Yan Yan. 2019. Multi-Channel Attention Selection GAN with Cascaded Semantic Guidance for Cross-View Image Translation. In *CVPR*.
Harrish Thasarathan and Mehran Ebrahimi. 2019. Artist-guided semiautomatic animation colorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.
Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. 2022. Clipasso: Semantically-aware object sketching. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–11.
Chao Wang, Haiyong Zheng, Zhibin Yu, Ziqiang Zheng, Zhaorui Gu, and Bing Zheng. 2018b. Discriminative region proposal adversarial networks for high-quality image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*. 770–785.
Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018a. High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision*

and Pattern Recognition (CVPR).

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

Holger Winnemöller. 2011. XDoG: Advanced Image Stylization with EXtended Difference-of-Gaussians. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering* (Vancouver, British Columbia, Canada) *(NPAR '11)*. Association for Computing Machinery, New York, NY, USA, 147–156. https://doi.org/10.1145/2024676.2024700

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Xiao Yang Yiheng Zhu Xiaohui Shen Xiaoyu Xiang, Ding Liu. 2021. Anime2Sketch: A Sketch Extractor for Anime Arts with Deep Networks. https://github.com/Mukosame/Anime2Sketch.

Shaoan Xie, Mingming Gong, Yanwu Xu, and Kun Zhang. 2021. Unaligned image-to-image translation by learning to reweight. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14174–14184.

Saining Xie and Zhuowen Tu. 2015. Holistically-Nested Edge Detection. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 1395–1403. https://doi.org/10.1109/ICCV.2015.164

Jiu Xu, Björn Stenger, Tommi Kerola, and Tony Tung. 2017. Pano2cad: Room layout from a single panorama image. In *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 354–362.

Xuemiao Xu, Minshan Xie, Peiqi Miao, Wei Qu, Wenpeng Xiao, Huaidong Zhang, Xueting Liu, and Tien-Tsin Wong. 2021. Perceptual-Aware Sketch Simplification Based on Integrated VGG Layers. *IEEE Transactions on Visualization and Computer Graphics* 27, 1 (2021), 178–189. https://doi.org/10.1109/TVCG.2019.2930512

Chuan Yan, David Vanderhaeghe, and Yotam Gingold. 2020. A Benchmark for Rough Sketch Cleanup. *ACM Transactions on Graphics (TOG)* 39, 6, Article 163 (Nov. 2020), 14 pages. https://doi.org/10.1145/3414685.3417784

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

Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. 2020. Unpaired portrait drawing generation via asymmetric cycle mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8217–8225.

Mingcheng Yuan and Edgar Simo-Serra. 2021. Line Art Colorization With Concatenated Spatial Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 3946–3950.

Lvmin Zhang, Chengze Li, Tien-Tsin Wong, Yi Ji, and Chunping Liu. 2018b. Two-stage sketch colorization. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–14.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018a. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 586–595. https://doi.org/10.1109/CVPR.2018.00068

Wenzhao Zheng, Jiwen Lu, and Jie Zhou. 2020. Structural deep metric learning for room layout estimation. In *European Conference on Computer Vision*. Springer, 735–751.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.

Changqing Zou, Haoran Mo, Chengying Gao, Ruofei Du, and Hongbo Fu. 2019. Language-based colorization of scene sketches. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–16.