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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. 0730-0301/2023/8-ART56 $15.00 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 [Illyasviel 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 [Ashhtari 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 50 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 [Illyasviel 2017], Anime2sketch [Xiaoou Yang 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 [Ashhtari 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 [Ashhtari 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,b] 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; Siaorhin 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 [Cí et al. 2018; Kim et al. 2019; Lee et al. 2020a; Liu et al. 2022; Ma et al. 2018b; Thasaratthan 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
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 is in the same domain as that of the reference input. Here, $C_i$ represents the input color image and $D$ represents the input sketch. $O$ is the output sketch domain generator to preserve visual similarity between unpaired data. In addition, 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 $E_c$. The details of the attentions are as follows.

For input feature maps, $E_c(C_i) \in \mathbb{R}^{C_i \times H \times W}$ and $E_c(R_i) \in \mathbb{R}^{C_i \times H \times W}$, 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 $CH_a \in \mathbb{R}^{C_i \times 1 \times 1}$) as follows:

$$SP_a(E_c(C_i)) = \sigma(f_{3 \times 3}([\text{AvgPool}^{3 \times 3}(E_c(C_i)); \text{MaxPool}^{3 \times 3}(E_c(C_i))]))$$

(1)

$$= \sigma(f_{3 \times 3}([E_c(C_i)]_{avg}; E_c(C_i)]_{max})),$$

$$CH_a(E_c(R_i)) = \sigma(\text{MLP}(\text{AvgPool}^{ch}(E_c(R_i))) + \text{MLP}(\text{MaxPool}^{ch}(E_c(R_i))))$$

(2)

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

![Diagram](image-url)
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 \( \sigma \) are used for both attentions. The feature sizes after the pooling layers are \( \text{AvgPool}^k, \text{MaxPool}^k \in \mathbb{R}^{C \times 1 \times 1} \) and \( \text{AvgPool}^{pP}, \text{MaxPool}^{pP} \in \mathbb{R}^{1 \times H \times W} \).

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

\[
\text{ADAIN}_a(E_a(C_i)) \odot \text{SP}_a(E_a(C_i)), E_r(R_i) \odot \text{CH}_a(E_r(R_i))).
\] (3)

These normalized features are then concatenated to the features from \( \text{ADAIN}_r \) 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.

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 [Ashi 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 \( O = G_x(C_i, R_i) \) generated by the sketch domain generator \( G_x \) 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:

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

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:

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

(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, \( \text{HED}(C_i) \) and \( \text{HED}(R_o) \), 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. \( \phi_l \) 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:

$$L_{CyC} = E_{C_i, R_o} [||C_i - R_o||_1]$$

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:

$$L_{adv} = E_{R_i} [\log(D(R_i))] + E_{O} [\log(1 - D(O))]$$

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

$$\min_G \max_D L_{total} = \lambda_{style} L_{style} + \lambda_{line} L_{line} + \lambda_{cyC} L_{cyC} + \lambda_{adv} L_{adv}$$

The parameters used in Eq. 8 are $\lambda_{style} = 5 - \frac{4.5i}{n}$, $\lambda_{cyC} = 10$, and $\lambda_{adv} = 1$, 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). 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 et al. 2018].

1CreativeCommons 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.

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.

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.

Table 3. Results of the cyclic evaluation

<table>
<thead>
<tr>
<th>Cyclic evaluation</th>
<th>PSNR↑</th>
<th>FID↓</th>
<th>LPIPS↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>35.84</td>
<td>91.97</td>
<td>0.1288</td>
</tr>
<tr>
<td>Ref2sketch</td>
<td>35.34</td>
<td>130.02</td>
<td>0.1643</td>
</tr>
<tr>
<td>Park et al. [2020b]</td>
<td>34.48</td>
<td>155.44</td>
<td>0.2824</td>
</tr>
<tr>
<td>MUNIT</td>
<td>33.92</td>
<td>151.04</td>
<td>0.2987</td>
</tr>
</tbody>
</table>

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 if 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.

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-colored 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.

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].

<table>
<thead>
<tr>
<th>Dataset</th>
<th>PSNR↑</th>
<th>FID↓</th>
<th>LPIPS↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>29.40</td>
<td>227.69</td>
<td>0.4802</td>
</tr>
<tr>
<td>Canny</td>
<td>28.68</td>
<td>268.10</td>
<td>0.5892</td>
</tr>
<tr>
<td>SketchKeras</td>
<td>27.99</td>
<td>283.69</td>
<td>0.6775</td>
</tr>
<tr>
<td>XDoG</td>
<td>28.55</td>
<td>239.63</td>
<td>0.5931</td>
</tr>
<tr>
<td>Simo-Serra et al. [2016]</td>
<td>28.49</td>
<td>270.99</td>
<td>0.6295</td>
</tr>
<tr>
<td>Combined</td>
<td>28.78</td>
<td>232.37</td>
<td>0.5604</td>
</tr>
</tbody>
</table>

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

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.

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


and Pattern Recognition (CVPR).


