Reference-Based Sketch Image Colorization using Augmented-Self Reference and Dense Semantic Correspondence

Junsoo Lee*,1, Eungyeup Kim*,1, Yunsung Lee2, Dongjun Kim1, Jaehyuk Chang3, Jaegul Choo1
1KAIST, 2Korea University, 3NAVER WEBTOON Corp.
{junsoolee93,eykim94,rassilon,jchoo}@kaist.ac.kr,
swack9751@korea.ac.kr, jaehyuk.chang@webtoonscorp.com

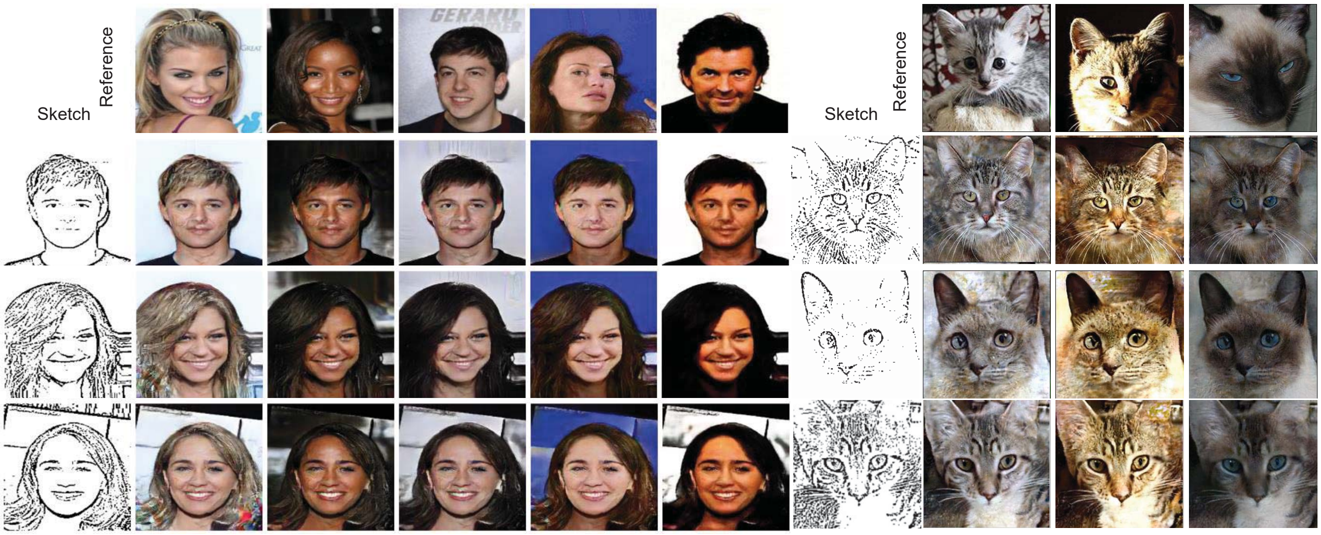


Figure 1: Qualitative results of our method on the CelebA [24] and ImageNet [31] dataset respectively. Each row has the same content while each column has the same reference.

Abstract

This paper tackles the automatic colorization task of a sketch image given an already-colored reference image. Colorizing a sketch image is in high demand in comics, animation, and other content creation applications, but it suffers from information scarcity of a sketch image. To address this, a reference image can render the colorization process in a reliable and user-driven manner. However, it is difficult to prepare for a training data set that has a sufficient amount of semantically meaningful pairs of images as well as the ground truth for a colored image reflecting a given reference (e.g., coloring a sketch of an originally blue car given a reference green car). To tackle this challenge, we propose to utilize the identical image with geometric distortion as a virtual reference, which makes it possible to secure the ground truth for a colored output image. Furthermore, it naturally provides the ground truth for dense semantic correspondence, which we utilize in our internal attention mechanism for color transfer from reference to sketch input. We demonstrate the effectiveness of our approach in various types of sketch image colorization via quantitative as well as qualitative evaluation against existing methods.

1. Introduction

Early colorization tasks [42, 18, 19] have focused on colorizing a grayscale image, which have shown great progress so far. More recently, the task of colorizing a given sketch or outline image has attracted a great deal of attention in both computer vision and graphics communities, due to its significant needs in practice. Compared to a grayscale im-

* indicates equal contribution

age, which still contains the pixel intensity, a sketch image is information-scarce, making its colorization challenging in nature. To remedy this issue, generally two types of approach of imposing additional conditions to the sketch image have been explored: user hints and reference image.

As explained in Section 2.2, there are previous works utilizing a reference or already-colored image, which shares the same semantic object of the target image. It requires an ability for the model to establish visual correspondences and inject colors through the mappings from the reference to the target. However, due to the huge information discrepancy between the sketch and reference, the sketch colorization guided by the reference is still under-explored compared to other sketch-based tasks (Section 2.1). Moreover, there are few datasets containing the labels of the correspondence between the two images, and the cost of generating a reliable matching of source and reference becomes a critical bottleneck for this task over a wide range of domains.

In this work, we utilize an augmented-self reference which is generated from the original image by both color perturbation and geometric distortion. This reference contains the most of the contents from original image itself, thereby providing a full information of correspondence for the sketch, which is also from the same original image. Afterward, our model explicitly transfers the contextual representations obtained from the reference into the spatially corresponding positions of the sketch by the attention-based pixel-wise feature transfer module, which we term the spatially corresponding feature transfer (SCFT) module. Integration of these two methods naturally reveals groundtruth spatial correspondence for directly supervising such an attention module via our similarity-based triplet loss. This direct supervision encourages the network to be fully optimized in an end-to-end manner from the scratch and does not require any manually-annotated labels of visual correspondence between source-reference pairs. Furthermore, we introduce an evaluation metric which measures how faithfully the model transfers the colors of the reference in the corresponding regions of sketch.

Both qualitative and quantitative experiments indicate that our approach exhibits the state-of-the-art performance to date in the task of information-scarce, sketch colorization based on a reference image. These promising results strongly demonstrate its significant potentials in practical applications in a wide range of domains.

2. Related work

2.1. Sketch-based Tasks

Sketch roughly visualizes the appearances of a scene or object by a series of lines. Thanks to its simple, easy-to-draw, and easy-to-edit advantages, sketch has been utilized in several tasks including image retrieval [17], sketch recognition [22], sketch generation [3, 26], and image inpainting [28]. However, due to the lack of texture and color information in sketch image, the research on sketch-based image colorization, especially reference-based colorization, is quite challenging and still under-explored.

2.2. Conditional Image Colorization

The automatic colorization has a limitation that users cannot manipulate the output with their desired color. To tackle this, recent methods come up with the idea of colorizing images with condition of the color given by users, such as scribbles [33], color palette [43, 25, 39], or text tags [15]. Even though these approaches have shown the impressive results in terms of the multi-modal colorization, they unavoidably require both precise color information and the geometric hints provided by users for every step.

To overcome the inconvenience, an alternative approach, which utilizes an already colored image as a reference, has been introduced. Due to the absence of geometric correspondence at the input level, early studies [14, 1, 23, 4, 6, 2] utilized low-level hand-crafted features to establish visual correspondence. Recent studies [8, 41, 35] compose the semantically close source-reference pairs by using features extracted from the pre-trained networks [8, 41] or color histogram [35] and exploit them in their training. These pair composition techniques however tend to be sensitive to domains, thereby limit their capability in a specific dataset.

Our work presents a novel training scheme to learn visual correspondence by generating augmented-self reference in the self-supervised manner at the training time, and then demonstrates it's scalability on various type of datasets.

3. Proposed method

In this section, we present our proposed model in detail, as illustrated in Fig. 2. We first describe overall workflow of the model and its two novel components called (1) Augmented-Self Reference Generation (Section 3.2) and (2) Spatially Corresponding Feature Transfer Module (Section 3.3). We then present our loss functions in detail.

3.1. Overall Workflow

As illustrated in Fig. 2, given a color image I in our dataset, we first convert it into its sketch image I_s using an outline extractor. Additionally, we generate an augmented-self reference image I_r by applying the thin plate splines (TPS) transformation. Taking these two images I_s and I_r as inputs, our model first encodes them into activation maps f_s and f_r using two independent encoders $E_s(I_s)$ and $E_r(I_r)$, respectively.

To transfer the information from I_r to I_s, we present a SCFT module inspired by a recently proposed self-attention mechanism [36], which computes dense correspondences

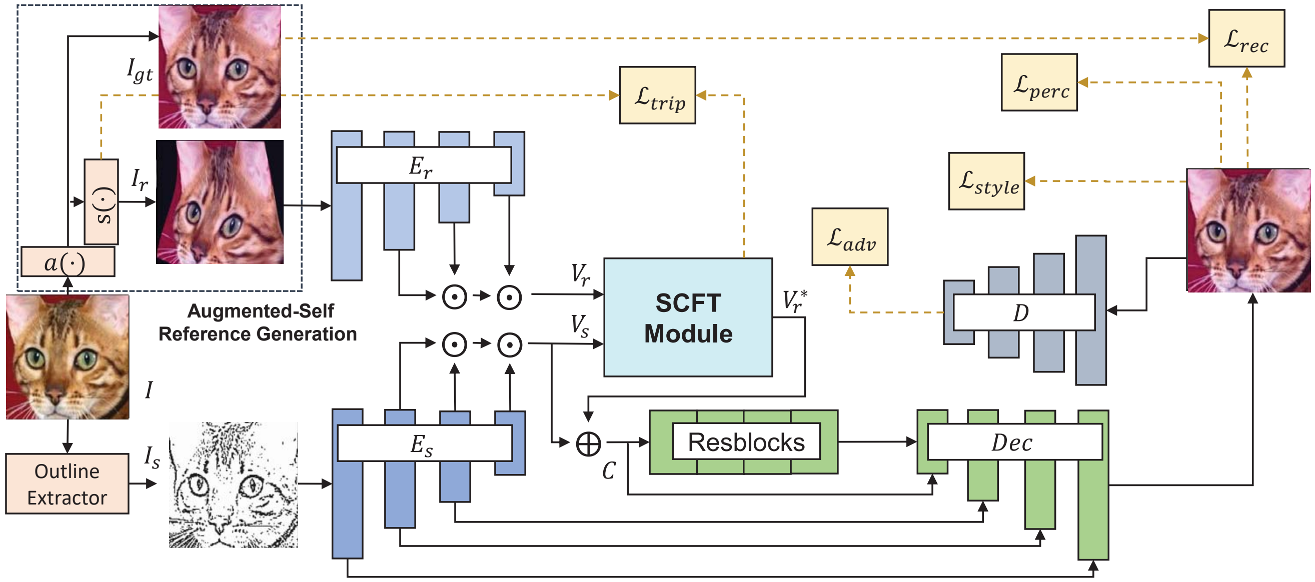


Figure 2: An overall workflow of our self-augmented learning process.

between every pixel pair of I_r to I_s. Based on the visual mappings from SCFT, context features fusing the information between I_r and I_s passes through several residual blocks and our U-net-based decoder [30] sequentially to obtain the final colored output.

3.2. Augmented-Self Reference Generation

To generate a reference color image I_r for a given sketch image I_s, we apply to original color image I two nontrivial transformations, appearance and spatial transformation. Since I_r is essentially generated from I, these processes guarantee that the useful information to colorize I_s exists in I_r, which encourages the model to reflect I_r in the colorization process. The details on how these transformations operate are described as follows. First, the appearance transformation $a(\cdot)$ adds a particular random noise per each of the RGB channel of I. The resulting output $a(I)$ is then used as the ground truth I_{gt} for the colorization output of our model. The reason why we impose color perturbation for making reference is to prevent our model from memorizing color bias, which means that a particular object is highly correlated with the single ground truth color in train data (i,e., a red color for apples). Given different reference in every iteration, our model should reconstruct different colored output for the same sketch, by leveraging I_r as the only path to restore I_{gt}. In other words, it encourages the model to actively utilize the information from E_r not just from E_s and generates reference-aware outputs at test time. Afterwards, we further apply the TPS transformation $s(\cdot)$, a non-linear spatial transformation operator to $a(I)$ (or I_{gt}), resulting in our final reference image I_r. This prevents our model from lazily bringing the color in the same pixel position from I_r, while enforcing our model to identify semantically meaningful spatial correspondences even for a reference image with a spatially different layout, e.g., different poses.

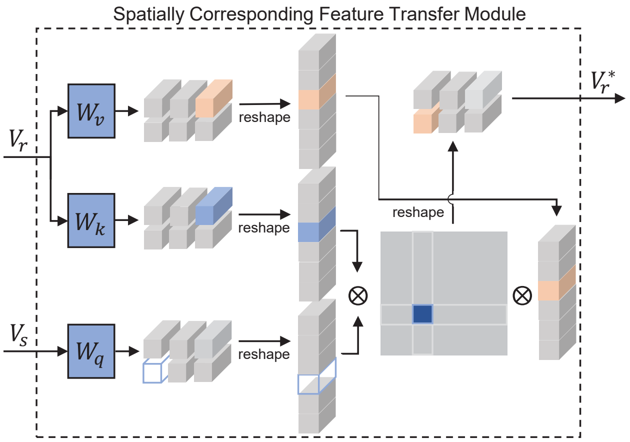


Figure 3: An illustration of spatially corresponding feature transfer (SCFT) module. SCFT establishes the dense correspondence mapping through attention mechanism.

3.3. Spatially Corresponding Feature Transfer

The goal of this module is to learn (1) which part of a reference image to bring the information from as well as (2) which part of a sketch image to transfer such information to, i.e., transferring the information from where to where. Once

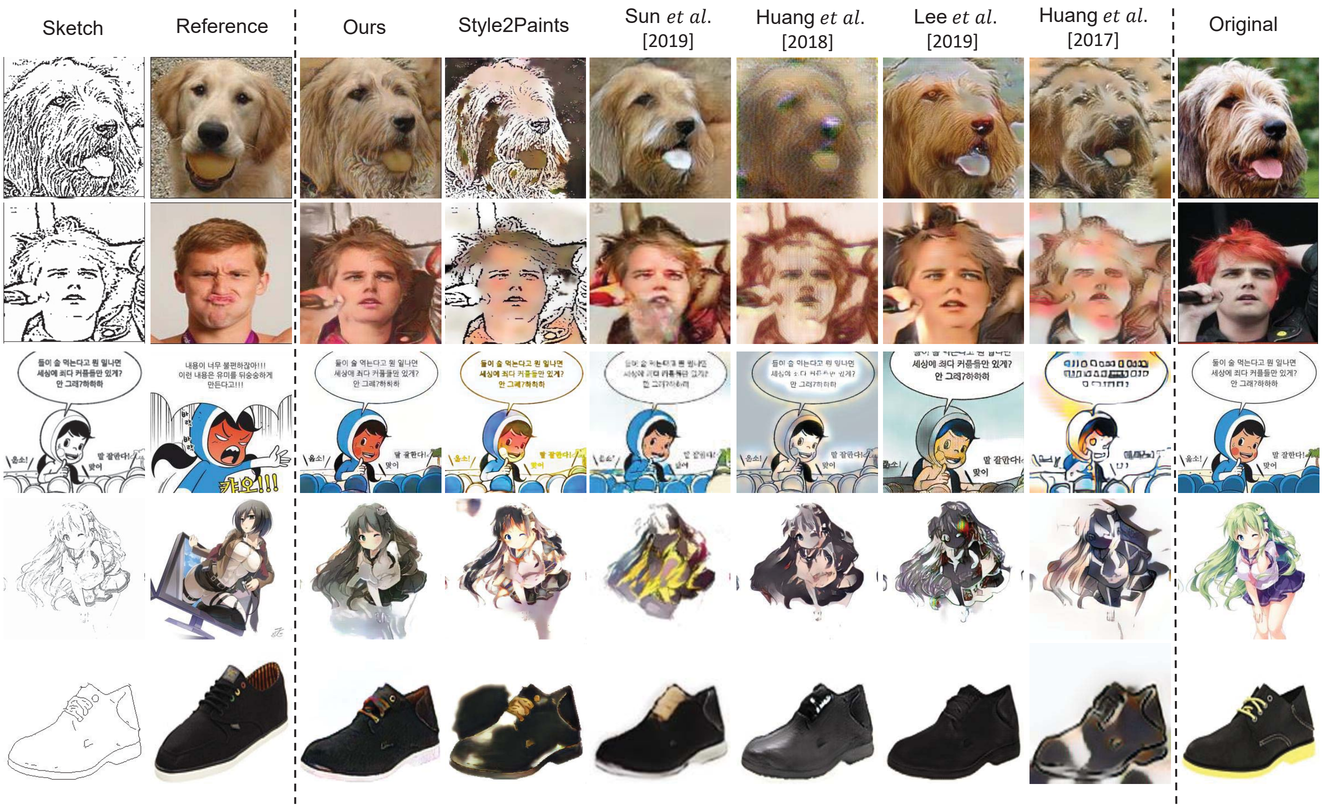


Figure 4: Qualitative comparison of colorize results with the baselines trained on the wide range of datasets. Note that the goal of our task does not reconstruct the original image. All results are generated from the unseen images. Please refer to the supplementary material for details.

obtaining this information as an attention map, our model transfers the feature information from a particular region of a reference to its semantically corresponding pixel of a sketch.

To begin with, each of the two encoders E_r and E_s consists of L convolutional layers, producing L activation maps $(f^1, f^2, \cdots, f^L)$ including intermediate outputs. Now we downsample each of them to match the spatial size of f^L and concatenate them along the channel dimensions, forming the final activation map V, i.e.,

$$V = \left[\varphi(f^1); \varphi(f^2); \cdots; f^{l_p}\right] \tag{1}$$

where φ denotes a spatially downsampling function of an input activation map $f^l \in \mathbb{R}^{h_l \times w_l \times c_l}$ to the size of $f^{l_p} \in \mathbb{R}^{h_p \times w_p \times c_p}$. ";" denotes the channel-wise concatenation operator. In this manner, we capture all the available low- to high-level features simultaneously.

Now we reshape V into $\bar{V} = [v_1, v_2, \cdots, v_{hw}] \in \mathbb{R}^{d_v \times hw}$, where $v_i \in \mathbb{R}^{d_v}$ indicates a feature representation of the i^{th} region of the given image and $d_v = \sum_{l=1}^{L} c_l$. We then obtain v_i^s of $\bar{V}_s$ and v_j^r of $\bar{V}_r$ from the outputs of the sketch encoder E_s and the reference encoder E_r, respectively. Given v_i^s and v_j^r, our model computes an attention matrix $\mathcal{A} \in \mathbb{R}^{hw \times hw}$ whose element α_{ij} is computed by the scaled dot product [36], followed by a softmax function within each row, i.e.,

$$\alpha_{ij} = \operatorname*{softmax}_j \left(\frac{(W_q v_i^s) \cdot (W_k v_j^r)}{\sqrt{d_v}} \right), \tag{2}$$

where $W_q, W_k \in \mathbb{R}^{d_v \times d_v}$ represent the linear transformation matrix into a query and a key vector, respectively, in the context of a self-attention module, and $\sqrt{d_v}$ represents a scaling factor. α_{ij} is a coefficient representing how much information v_i^s should bring from v_j^r. Now we can obtain the context vector v_i^* of the position i as

$$v_i^* = \sum_j \alpha_{ij} W_v v^{r_j}, \tag{3}$$

where $W_v \in \mathbb{R}^{d_v \times d_v}$ is the linear transformation matrix into a value vector containing the color feature in a semantically related region of a reference image.

Finally, v_i^* is added to the original feature v_i^s of a sketch image to form the feature vector enriched by the information of the corresponding region in the reference image, i.e.,

$$c_i = v_i^s + v_i^* \tag{4}$$

c_i is then fed into the decoder to synthesize a colored image.

3.4. Objective Functions

Similarity-Based Triplet Loss. When applying the spatial transformation $s(\cdot)$, each pixel value in the output image is represented as a weighted average of pixels in the input image, revealing the spatial correspondences of pixel pairs between I_s and I_r. In other words, we can obtain the full information of the weight w_{ij}, which represents how much the i^{th} pixel position of the input image, or a query, is related to the j^{th} pixel position of the output, or a key. Then, the value of w_{ij} can be considered as the pixel-to-pixel correspondence, which can work as the groundtruth for supervising how semantically related the pixel of the reference to a particular pixel of sketch image.

Utilizing this pixel-level correspondence information, we propose a similarity-based triplet loss, which is a variant of triplet loss [34], to directly supervise the affinity between the pixel-wise query and key vectors used to compute the attention map $\mathcal{A}$ in Eq. (2). The proposed loss term is computed as

$$\mathcal{L}_{tr} = \max(0, [-S(v_q, v_k^p) + S(v_q, v_k^n) + \gamma]), \tag{5}$$

where $S(\cdot,\cdot)$ computes the scaled dot product. Given a query vector v_q as an anchor, v_k^p indicates a feature vector sampled from the positive region, and v_k^n is a negative sample. γ denotes a margin, which is the minimum distance $S(v_q, v_k^p)$ and $S(v_q, v_k^n)$ should maintain. $\mathcal{L}_{tr}$ encourages the query representation to be close to the correct (positive) key representation, while penalizing to be far from the wrong (negatively sampled) one. This loss plays a crucial role in directly enforcing our model to find the semantically matching pairs and reflect the reference color into the corresponding position.

The reason we adopt triplet loss instead of commonly used losses such as L_1-loss is that the latter can overly penalize the affinities between semantically close but spatially distant query and key pixel pairs. This misleading result can be mitigated by only penalizing two cases: the semantically closest pair (positive sample) and randomly-sampled except it (negative sample), which is basically a triplet loss.

We further conduct a user study to compare the effects of our triplet loss to another possible loss, i.e., L_1-loss and no supervision. Details about the experimental settings and results are explained in Section 6.2 in the supplementary material.

L1 Loss. Since the groundtruth image I_{gt} is generated as Section 3.2, we can directly impose a reconstruction loss to penalize the network for the color difference between the output and the ground truth image as below:

$$\mathcal{L}_{rec} = \mathbb{E}\left[\| G(I_s, I_r) - I_{gt} \|_1\right]. \tag{6}$$

Adversarial Loss. The discriminator D, as an adversary of the generator, has an objective to distinguish the generated images from the real ones. The output of real/fake classifier $D(X)$ denotes the probability of an arbitrary image X to be a real one. We adopt *conditional GANs* which use both a generated sample and additional conditions [29, 38, 12]. In this work, we leverage the input image I_s as a condition for the adversarial loss since it is important to preserve the content of I_s as well as to generate a realistic fake image. The loss for optimizing D is formulated as a standard cross-entropy loss as

$$\begin{aligned}\mathcal{L}_{adv} = \mathbb{E}_{I_{gt}, I_s}\left[\log D(I_{gt}, I_s)\right] \\ + \mathbb{E}_{I_s, I_r}\left[\log(1 - D(G(I_s, I_r), I_s))\right].\end{aligned} \tag{7}$$

Perceptual Loss. As shown in previous work [28], perceptual loss [13] encourages a network to produce an output that is perceptually plausible. This loss penalizes the model to decrease the semantic gap, which means the difference of intermediate activation maps between the generated output $\hat{I}$ and the ground truth I_{gt} from the ImageNet [31] pretrained network. We employ a perceptual loss using multilayer activation maps to reflect not only high-level semantics but also low-level styles as

$$\mathcal{L}_{perc} = \mathbb{E}\left[\sum_l \| \phi_l(\hat{I}) - \phi_l(I_{gt}) \|_{1,1}\right], \tag{8}$$

where ϕ_l represents the activation map of the l'th layer extracted at the *relu*_1 from the VGG19 network.

Style Loss. Sajjadi *et al.* [32] has shown that the style loss which narrow the difference between the covariances of activation maps is helpful for addressing checkerboard artifacts. Given $\phi_l \in \mathbb{R}^{C_l \times H_l \times W_l}$, the style loss is computed as

$$\mathcal{L}_{style} = \mathbb{E}\left[\| \mathcal{G}(\phi_l(\hat{I})) - \mathcal{G}(\phi_l(I_{gt})) \|_{1,1}\right], \tag{9}$$

where $\mathcal{G}$ is a gram matrix.

In summary, the overall loss function for the generator G and discriminator D is defined as

$$\begin{aligned}\min_G \max_D \mathcal{L}_{total} = \lambda_{tr}\mathcal{L}_{tr} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv} \\ + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{style}\mathcal{L}_{style}.\end{aligned} \tag{10}$$

3.5. Implementation Details

We implement our model with the size of input image fixed in 256×256 on every datasets. For training, we set the coefficients for each loss functions as follows: $\lambda_{adv} = 1$, $\lambda_{rec} = 30$, $\lambda_{tr} = 1$, $\lambda_{perc} = 0.01$, and $\lambda_{style} = 50$. We

	ImageNet			Human Face	Comics		Hand-drawn
Methods	Cat	Dog	Car	CelebA	Tag2pix	Yumi's Cells	Edges→Shoes
Sun *et al.* [35]	160.65	168	192.00	75.66	122.14	72.45	124.98
Huang *et al.* [11]	281.44	271.47	258.36	173.12	76.00	132.90	86.43
Lee *et al.* [21]	151.52	172.22	70.07	68.43	91.65	63.34	109.29
Huang *et al.* [10]	257.39	268.69	165.84	160.22	97.40	148.52	190.16
(a) Ours w/o $\mathcal{L}_{tr}$	77.39	109.49	54.07	53.58	47.68	51.34	79.85
(b) Ours full	**74.12**	**102.83**	**52.23**	**47.15**	**45.34**	**49.29**	**78.32**

Table 1: Quantitative comparisons over the datasets with existing baselines by measuring FID [9] score: a lower score is better.

set the margin of the triplet loss $\gamma = 12$ for overall data. We use Adam solver [16] for optimization with $\beta_1 = 0.5$, $\beta_2 = 0.999$. The learning rate of generator and discriminator are initially set to 0.0001 and 0.0002 for each. The detailed network architectures are described in Section 6.5 of supplementary material.

4. Experiments

This section demonstrates the superiority of our approach on wide range of domain datasets (Section 4.1) including real photos, human face and anime (comics). We newly present an evaluation metric, named SC-PSNR described in Section 4.2, to measure the faithfulness of reflecting the style of the reference. Afterwards, we compare our method against the several baselines of related tasks quantitatively as well as qualitatively (Section 4.3). An in-depth analysis of our approach is described across Section 4.4-4.5.

4.1. Datasets

Tag2pix Dataset. We use Tag2pix dataset [15], which contains large-scale anime illustrations filtered from Danbooru2017 [7], to train our model for comic domain. Although there are various tag labels on this dataset, we only utilize images to train the model owing to our self-supervised training scheme. It consists of one character object with white background images. We partition into 54,317 images for train, 6036 images for test and then combine source-reference pairs by randomly sampled from the test set for evaluation.

Yumi Dataset. Like Yoo *et al.* [40], we collect images from the online cartoon named *Yumi's Cells* for the outline colorization of the anime domain. The dataset contains repeatedly emerging characters across 329 episodes. With this limited variety of characters, the network is required to find the correct character matching even if there is no explicit character supervisions. We randomly split into a train set of 7,014 images and test set of 380 images, and then manually construct source-reference pairs from the testset to evaluate the performance of the models.

	SC-PSNR (dB)		
Methods	Cat	Dog	Car
Sun *et al.* [35]	9.65	11.19	9.42
Huang *et al.* [11]	10.33	12.67	8.45
Lee *et al.* [21]	11.54	12.08	9.94
Huang *et al.* [10]	9.25	9.49	7.77
(a) Ours w/o $\mathcal{L}_{tr}$	12.76	13.73	10.56
(b) Ours full	**13.23**	**14.37**	**11.34**

Table 2: Quantitative comparisons over the SPair-71k with existing baselines by measuring SC-PSNR (dB) score: a higher score is better.

SPair-71k Dataset. SPair-71k dataset [27], which is manually annotated for a semantic correspondence task, consists of total 70,958 pairs of images from PASCAL 3D+ [37] and PASCAL VOC 2012 [5]. We select two non-rigid categories (cat, dog) and one rigid category (car), of which we can gather sufficient data points from ImageNet [31]. Note that this dataset is used to measure SC-PSNR (Section. 4.2) score only for the evaluation purpose.

ImageNet Dataset. As above-mentioned, we collect subclasses that correspond to three categories (i.e., cat, dog, car) from ImageNet [31] dataset and use them for training data. Images in each class are randomly divided into two splits with an approximate ratio of 9:1 for training and validation.

Human Face Dataset. Our method can be applied to colorize a sketch image of human face domain as well. To support this claim, we leverage CelebA [24] dataset, which have commonly been used for image-to-image translation or style transfer tasks. Training and validation sets are composed as the ImageNet dataset are.

Edges→Shoes Dataset. We use Edges→Shoes dataset, which contains pairs of sketch-color shoes images that have been widely used in image-to-image translation tasks [20, 11] as well. This enables a valid evaluation between our method and existing unpaired image-to-image translation

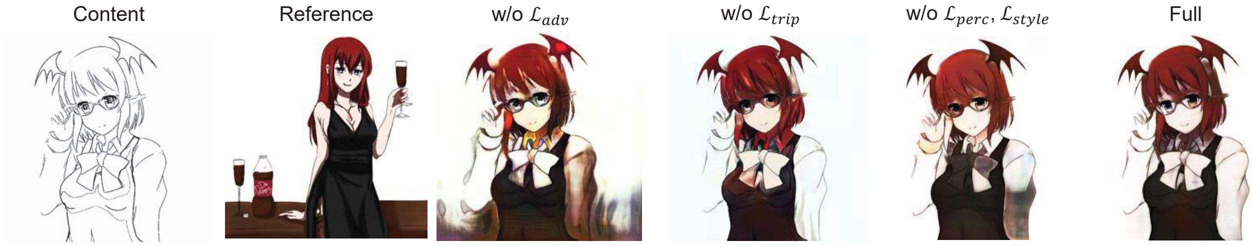


Figure 5: A qualitative example presenting the effectiveness of different loss functions.

	ImageNet			Human Face	Comics		Hand-drawn
Loss Functions	Cat	Dog	Car	CelebA	Tag2pix	Yumi's Cells	Edges→Shoes
$\mathcal{L}_{rec}$	82.10	143.76	68.45	77.70	58.00	52.86	91.10
$\mathcal{L}_{rec} + \mathcal{L}_{adv}$	78.56	110.86	56.54	54.75	48.71	51.96	82.55
$\mathcal{L}_{rec} + \mathcal{L}_{adv} + \mathcal{L}_{perc} + \mathcal{L}_{style}$	77.39	109.49	54.07	53.58	47.68	51.34	79.85
$\mathcal{L}_{rec} + \mathcal{L}_{adv} + \mathcal{L}_{perc} + \mathcal{L}_{style} + \mathcal{L}_{tr}$	**74.12**	**102.83**	**52.23**	**47.15**	**45.34**	**49.29**	**78.32**

Table 3: FID scores [9] according to the ablation of loss function terms described in Section 4.4. A lower score is better.

approaches.

4.2. Evaluation Metrics

Semantically Corresponding PSNR. This work proposes a novel evaluation metric to measure how faithfully the model transfers the style of reference in the corresponding regions. In the traditional automatic colorization setting where a groundtruth image is available, pixel-level evaluation metric, such as peak signal-to-noise ratio (PSNR), has been widely used. In reference-based colorization setting, however, there is no ground truth that have both the shape of the content and the style of the reference.

The key idea behind the semantically corresponding PSNR (SC-PSNR) is leveraging the datasets created for keypoint alignment tasks [5, 37, 27], thereby providing patch-level groundtruth. We use SPair-71k dataset [27] which contains semantically corresponding annotation pairs between two different images. Only the pixel values in a certain size of patch surrounding the corresponding keypoints of two images are used instead of the whole pixels for computing mean square error (MSE), and then PSNR is computed with the MSE. We refer to this measurement as the SC-PSNR.

Fig. 6 shows first and last two examples of images queried by the leftmost image. The list of images are retrieved in a decreasing order of the value of SC-PSNR being computed with query. This figures demonstrates that this metric captures perceptually plausible distance of the pixel values between the keypoint regions of two images.

***Frećhet Inception Distance* (FID) [9].** FID is a well-known metric for evaluating the performance of a generative model by measuring the Wasserstein-2 distance between the feature space representations of the real images and its generated outputs. A low score of FID indicates that the model generates the images with quality and diversity close to real data distribution.

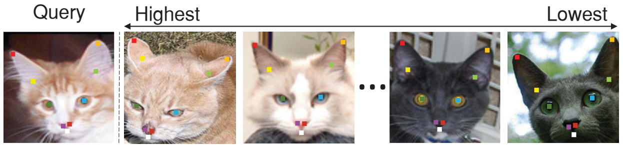


Figure 6: Different colors of points denote different keypoint annotations on cat face, e.g., eyes and noses.

4.3. Comparisons to Baselines

We compare our method against recent deep learning-based approaches on the various types of datasets both qualitatively and quantitatively. The baselines are selected from not only the colorization task [25, 35] but also the related problems tackling multi-modal image generation, such as exemplar-guided image translation [11, 21] and style transfer [10].

Fig. 4 shows the overall qualitative results of our model and other baselines on 5 different datasets. Datasets vary from real image domain like ImageNet or Human face dataset to sketch image domain like Edges→Shoes, Yumi's Cells, and Tag2pix. The leftmost and second column are sketch and reference, respectively. On every dataset our model brings the exact colors from the reference image and injects them into the corresponding position in the sketch.

For example, our model colorizes the character's face in third row with red color from the reference, while baselines tend not to fully transfer it. Likewise, in fifth row, inner side of the shoes and shoe sole are elaborately filled with the color exactly referencing the exemplar image.

We report on Table 1 the FID score calculated over the 7 different datasets. Our method outperforms the existing baselines by a large margin, demonstrating that our method has the robust capability of generating realistic and diverse images. Improved scores of our model with triplet loss indicates that $\mathcal{L}_{tr}$ plays a beneficial role in generating realistic images by directly supervising semantic correspondence.

Table 2 presents the other quantitative comparisons in regard to the SC-PSNR scores as described in Section 4.2. We measure SC-PSNR only over cat, dog and car dataset which are subclasses belonging to both ImageNet and SPair-71k [27]. Our method outperforms all the baseline models, demonstrating that our model is superior at establishing visual correspondences, and then generating suitable colors.

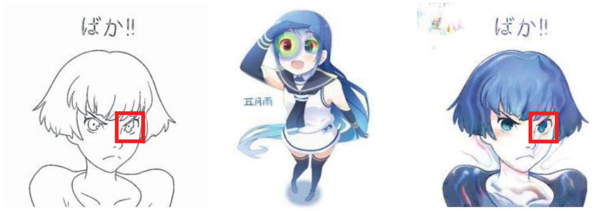
We conduct a user study for human evaluation on our model and other existing baselines, as shown in Fig. 8. The detailed experimental setting is described in Section 6.2 in the supplementary material. Our model occupies a large percentage of Top1 and Top2 votes, indicating that our method better reflects the color from the reference and generates more realistic outputs than other baselines.

4.4. Analysis of Loss Functions

We ablate the loss functions individually to analyze the effects of the functions qualitatively, as shown in Fig. 5 and quantitatively, as shown in Table 3. When we remove $\mathcal{L}_{adv}$, output image contains inaccurate colors emerging in the background and dramatically appears unrealistic. Without $\mathcal{L}_{tr}$, character's back hair, forehead and ribbon tail are colorized with wrong color or even not colorized. The FID score in Table 3 third row also represents that model generates unrealistic output. This degraded performance is due to the absence of supervision which encourages to match the semantically close regions between content and reference. When we remove $\mathcal{L}_{perc}$ and $\mathcal{L}_{style}$, the colorization tends to produce color bleeding or visual artifacts since there is no constraint to penalize the model for the semantic difference between the model output and the ground truth. Image generated with full losses have exact colors in its corresponding regions with fewer artifacts.

4.5. Visualization of Attention Maps

Fig. 7 shows an example of an attention map $\mathcal{A}$ learned by our SCFT module. In this module, each pixel from the sketch is used as a query to retrieve the relevant local information from the reference. In the case of left-eye region as a query (red square in (a)), we visualize the top three, highly-attentive regions in the reference image (a highlighted region in (b)). Based on this attention pattern, our model properly colorizes the left eye of a person in a sketch image (c) with blue color. For additional examples of visualizing attention maps for different sketch and reference images, we strongly encourage the readers to check out the Fig. 14 in the supplementary material for details.



(a) Sketch (b) Reference (c) Synthesized image

Figure 7: Visualization of our attention mechanism.

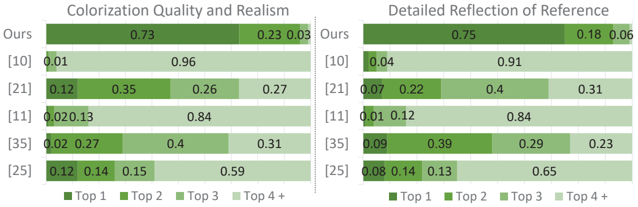


Figure 8: User study results. Percentage values are averaged over every datasets we experimented. Individual results are presented in Section 6.2 in supplementary material.

5. Conclusions

This paper presents a novel training scheme, integrating the augmented-self reference and the attention-based feature transfer module to directly learn the semantic correspondence for the reference-based sketch colorization task. Evaluation results demonstrate that our SCFT module exhibits the state-of-the-art performance over the diverse datasets, which demonstrates the significant potentials in practice. Finally, SC-PSNR, a proposed evaluation metric, effectively measures how the model faithfully reflects the style of the exemplar.

Acknowledgements. This work was partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)), the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (No. NRF-2019R1A2C4070420), and the National Supercomputing Center with supercomputing resources including technical support (KSC-2019-CRE-0133). Finally, we thank all researchers at NAVER WEBTOON Corp.

References

[1] Aurélie Bugeau, Vinh-Thong Ta, and Nicolas Papadakis. Variational exemplar-based image colorization. *IEEE Transactions on Image Processing*, 23(1):298–307, 2013. 2

[2] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf. Automatic image colorization via multimodal predictions. In *ECCV*, pages 126–139, 2008. 2

[3] Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *CVPR*, pages 9416–9425, 2018. 2

[4] Alex Yong-Sang Chia, Shaojie Zhuo, Raj Kumar Gupta, Yu-Wing Tai, Siu-Yeung Cho, Ping Tan, and Stephen Lin. Semantic colorization with internet images. *TOG*, 30(6):156:1–156:8, 2011. 2

[5] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 6, 7

[6] Raj Kumar Gupta, Alex Yong-Sang Chia, Deepu Rajan, Ee Sin Ng, and Huang Zhiyong. Image colorization using similar images. In *MM*, pages 369–378, 2012. 2

[7] Aaron Gokaslan Gwern Branwen. Danbooru2017: A large-scale crowdsourced and tagged anime illustration dataset. `https://www.gwern.net/Danbooru2017`, 2018. [Online; accessed 22-03-2018]. 6

[8] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. Deep exemplar-based colorization. *TOG*, 37(4):47, 2018. 2

[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. In *NIPS*, pages 6626–6637, 2017. 6, 7

[10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 6, 7

[11] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 172–189, 2018. 6, 7

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 5

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

[14] Hemant B Kekre and Sudeep D Thepade. Color traits transfer to grayscale images. In *2008 First International Conference on Emerging Trends in Engineering and Technology*, pages 82–85, 2008. 2

[15] Hyunsu Kim, Ho Young Jhoo, Eunhyeok Park, and Sungjoo Yoo. Tag2pix: Line art colorization using text tag with secat and changing loss. In *ICCV*, pages 9056–9065, 2019. 2, 6

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[17] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018. 2

[18] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *ECCV*, pages 577–593, 2016. 1

[19] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, pages 6874–6883, 2017. 1

[20] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, pages 35–51, 2018. 6

[21] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 2020. 6, 7

[22] Fang Liu, Xiaoming Deng, Yu-Kun Lai, Yong-Jin Liu, Cuixia Ma, and Hongan Wang. Sketchgan: Joint sketch completion and recognition with generative adversarial network. In *CVPR*, pages 5830–5839, 2019. 2

[23] Xiaopei Liu, Liang Wan, Yingge Qu, Tien-Tsin Wong, Stephen Lin, Chi-Sing Leung, and Pheng-Ann Heng. Intrinsic colorization. *TOG*, 27(5):152:1–152:9, 2008. 2

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 1, 6

[25] lllyasviel. style2paints. `https://github.com/lllyasviel/style2paints`, 2018. [Online; accessed 22-03-2018]. 2, 7

[26] Yongyi Lu, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. Image generation from sketch constraint using contextual gan. In *ECCV*, pages 205–220, 2018. 2

[27] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019. 6, 7, 8

[28] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 5

[29] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, pages 2642–2651. JMLR. org, 2017. 5

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[31] 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. 1, 5, 6

[32] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, pages 4491–4500, 2017. 5

[33] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *CVPR*, pages 5400–5409, 2017. 2

[34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 5

[35] Tsai-Ho Sun, Chien-Hsun Lai, Sai-Keung Wong, and Yu-Shuen Wang. Adversarial colorization of icons based on contour and color conditions. In *MM*, pages 683–691, 2019. 2, 6, 7

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 2, 4

[37] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *2014 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 75–82, 2014. 6, 7

[38] 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 *CVPR*, pages 1316–1324, 2018. 5

[39] Taizan Yonetsuji. Paintschainer. `https://paintschainer.preferred.tech/index_en.html`, 2017. [Online; Accessed 22-03-2018]. 2

[40] Seungjoo Yoo, Hyojin Bahng, Sunghyo Chung, Junsoo Lee, Jaehyuk Chang, and Jaegul Choo. Coloring with limited data: Few-shot colorization via memory augmented networks. In *CVPR*, pages 11283–11292, 2019. 6

[41] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *CVPR*, pages 8052–8061, 2019. 2

[42] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, pages 649–666, 2016. 1

[43] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros. Real-time user-guided image colorization with learned deep priors. *TOG*, 36(4), 2017. 2