mathematics

MDPI

Article

Exemplar-Based Sketch Colorization with Cross-Domain Dense Semantic Correspondence

Jinrong Cui *, Haowei Zhong, Hailong Liu and Yulu Fu

College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China; zhw233@stu.scau.edu.cn (H.Z.); hailong1394793403@stu.scau.edu.cn (H.L.); fuyulu@stu.scau.edu.cn (Y.F.)
* Correspondence: jinrongcui@scau.edu.cn

Abstract: This paper aims to solve the task of coloring a sketch image given a ready-colored exemplar image. Conventional exemplar-based colorization methods tend to transfer styles from reference images to grayscale images by employing image analogy techniques or establishing semantic correspondences. However, their practical capabilities are limited when semantic correspondences are elusive. This is the case with coloring for sketches (where semantic correspondences are challenging to find) since it contains only edge information of the object and usually contains much noise. To address this, we present a framework for exemplar-based sketch colorization tasks that synthesizes colored images from sketch input and reference input in a distinct domain. Generally, we jointly proposed our domain alignment network, where the dense semantic correspondence can be established, with a simple but valuable adversarial strategy, that we term the structural and colorific conditions. Furthermore, we proposed to utilize a self-attention mechanism for style transfer from exemplar to sketch. It facilitates the establishment of dense semantic correspondence, which we term the spatially corresponding semantic transfer module. We demonstrate the effectiveness of our proposed method in several sketch-related translation tasks via quantitative and qualitative evaluation.

Keywords: sketch colorization; image synthesis; reference-based colorization

MSC: 68T01

Citation: Cui, J.; Zhong, H.; Liu, H.; Fu, Y. Exemplar-Based Sketch Colorization with Cross-Domain Dense Semantic Correspondence. *Mathematics* **2022**, *10*, 1988. https://doi.org/10.3390/math10121988

Academic Editor: Ezequiel López-Rubio

Received: 14 April 2022
Accepted: 3 June 2022
Published: 9 June 2022

1. Introduction

Sketch roughly describes the attributes and appearances of an object or a scene by a series of lines, and sketch colorization, which assigns colors to binary line images to improve their visual quality while preserving the original semantic information. Nowadays, neural style translation has succeeded in image translation, which renders the image and changes its color and texture while keeping its content characteristics unchanged [1–6]. The previous neural translation methods perform well in grayscale images, but not in the conversion of sketch manuscript images. Therefore, the translation task on sketches has attracted a great deal of attention in both the content industry and computer vision. In contrast to the coloring task of sketch images, the grayscale coloring task is mainly based on the assumption that neighboring pixels with similar intensities in grayscale should have similar colors. Sketch images are information-scarce, making their colorization tasks naturally challenging. We consider that the previous method may fail to learn a more challenging mapping from sketches with intricate edges to colored images. Two types of methods of sketch colorization tasks have been explored: hint-based (e.g., strokes, palette, and text) approach and reference-based approach.

It comes up with an intuitive way to colorize a sketch with a small amount of auxiliary information given by users, such as stroke hint [7–10], color palette [11,12], and text label [13–15]. Although these hint-based colorization methods show impressive results, they still suffer from the requirement of unambiguous color information and precise spatial user inputs for every step. Therefore, a more convenient coloring mode appears, utilizing

exemplar images for sketch colorization. In the practice of exemplar colorization, a critical point is the preparation of a sufficiently large number of semantic training image pairs and the ground truth that reflects the color results of a given exemplar. One attempt [16] used geometric distortion and color perturbation to synthesize a pseudo ground truth. However, it suffers from the problems that failed to handle cross-domain samples well and easy to mode collapse. Therefore, some research is aimed at cross-domain learning and has been successfully employed in image translation. Early methods [17–22] focus on utilizing the low-level features to compose colorization. Although the above early methods broaden the thinking of style transfer, there are still many limitations: (1) The source image and target image are required to have a certain similarity in form and shape; (2) there are some deficiencies in the display of the global semantic features of the image; and (3) the style of the generated image is monotonous, and the texture diversity is not rich enough. To surmount such problems, recent studies [23–30] have explored the establishment of cross-domain correspondence between the exemplar and source input. An extension of Image Analogies [28] and Deep Analogy [29] tries to establish the dense semantically-meaningful correspondence of an input pair using pre-trained VGG layers. We deem that such methods may fail to handle sketch colorization. In order to consider the sketch (or mask, edge) format in the task of image translation, some studies [24,31,32] explicitly divide the exemplars into semantic regions and learn to synthesize different regions separately. Some research [23,27,30] utilizes the deep network for composing semantically close source-reference pairs or takes advantage of histograms [30] to exploit sketches in their training. In this manner, it managed to produce high-quality results. However, these methods are domain-specific and are unsuitable for sketch colorization with only complex edges composition. Additionally, the style only marries the global context style, regardless of spatially relevant information and partial local style.

Our concern is how to establish the dense correspondence between sketch and exemplar in a more efficient manner. Our motivations are mainly on two issues: Firstly, how to model and extract local and non-local styles from exemplar images more efficiently? Secondly, how to learn the mapping with desired style information extracted from exemplars while preserving the semantically-meaningful sketch composition. For the first issue, we proposed a cross-domain alignment module that transforms distinct domain inputs into a shared, embedded space to ulteriorly learn the dense correspondence in both local and non-local style manners. For the second case, we propose a module that explicitly transfers the canonical contextual representation to the spatial location of the sketch input through a self-attentive pixelated feature transfer mechanism, which we term the cross-domain spatially feature transfer module (CSFT). Finally, a set of spatially-invariant de-normalization blocks with a Moment Shortcut (MS) connection [33] are employed to synthesize the output progressively; then, a specific adversarial framework for colorization tasks, dual multiscale discriminators with the capability of distinguishing structural composition and style coloration, respectively, has been introduced in this paper to facilitate the joint training of alignment module and guide the reconstruction of stylized output. This indirect supervision departs from the requirement of manually-annotated samples with visual correspondence between source-exemplar pairs. It encourages the network can be fully optimized in an end-to-end manner.

Qualitative and quantitative experimental results show that our method outperforms previous methods and exhibits state-of-the-art performance. These promising results extensively demonstrate its great potential for practical applications in various fields. The main contributions of this paper can be summarized as follows:

- The cross-domain alignment module is proposed for imposing the distinct domain to a shared, embedded space for progressively aligning and outputting the warped image in a coarse-to-fine manner.
- To facilitate the establishment of dense correspondence, we proposed an explicit style transfer module utilizing self attention-based pixel-wise feature transfer mechanism, which we term the cross-domain spatially feature transfer module (CSFT).

- We proposed a specific adversarial strategy for exemplar-based sketch colorization to facilitate the imaging quality and stabilize the adversarial training.

2. Related Work

2.1. Image-to-Image Translation

Image-to-image translation is the problem of converting a possible representation of one scene into another, such as mapping a semantical mask to an RGB image or vice versa. Most previous prominent approaches show their ability on translation tasks with a generative adversarial network [34] that leverages either paired data [6,35,36] or unpaired data [37–39]. The previous generative models solve the image-to-image translation with different domains. However, they can only learn the latent representation between two specific different domains at a time, which makes it hard to deal with the transformation between multiple domains. Therefore, Liu et al. [40] designed the UNIT network based on GAN and VAE, and they realized the conversion from unsupervised image to image by learning a shared latent space. Then, Choi et al. [41] proposed starGAN, which is trained on multiple cross-domain datasets to realize multi-domain transformation. However, none of these methods concern the geometric gap between source content and style target. Additionally, previous methods ignore the capability of delicate control of the final output because the latent space representation is rather complex and implicit in correspondence of the exemplar style. In contrast, our cross-domain alignment module supports customization of final colorization results by a given user-guided exemplar in a coarse-to-fine manner of warping and refining, allowing users to control their designed effect flexibly.

2.2. Sketch-Based Tasks

A sketch is a rough visual representation of a scene or object by a set of lines and edges. It has been utilized in several computer vision tasks such as image retrieval [42,43], sketch generation [44,45], and sketch recognition [46]. Unlike other image-to-image translation methods, sketch colorization plays a unique role in content creation. Frans [47] used a user-defined color scheme colorization model based on GANs, but it hardly generated agreeable results. Ci et al. [7] explored the line art colorization in the field of animation by introducing ResNeXt and a pre-trained model to alleviate the problem of overfitting. Hati et al. [9] is based on Ci's model, introducing a double generator to improve visual fidelity but greatly increase the number of parameters. Style2Paints [8] was published as a famous project on Github with 14k stars, and the newest version is Style2Paints V4.5 beta. The V4.5 version can generate visually pleasing line art colorization results by splitting line art images into different parts and colorize them respectively. Zhang et al. [48] used U-Net residual architecture and an auxiliary classifier to preliminarily realize the animation style colorization tasks of sketches. Although these methods show impressive results for sketch-based coloring, they inevitably require precise color information and a certain amount of geometric cueing information that the user needs to provide at each step.

An alternative approach, which utilizes an already colored image as an exemplar to colorize sketches, has been introduced to surmount these inconveniences. Lee et al. [16] explored geometric augmented-self reference in the training process to generate forged sample pairs. Sun et al. [30] composed the semantically-related reference-pairs by color histogram. Lian et al. [49] explored an anime sketch colorization net without encoder using Spatially-Adaptive Normalization. However, these pair composition methods tend to be sensitive to domains, limiting their capability in a specific dataset. In contrast, our cross-domain model can be better applied to cross-domain learning and different types of datasets. At the same time, we have designed a novel adversarial strategy for sketch colorization to facilitate the final imaging quality.

2.3. Exemplar-Based Image Synthesis

More recently, researchers [25,50,51] have proposed to synthesize images from the semantic layout of the input under the guidance of exemplars. Zhang et al. [27] designs a novel end-to-end dual branch network architecture. When reliable reference pictures are

not available, it learns reasonable local coloring to generate meaningful reference pictures and makes a reasonable color prediction. Huang et al. [51] and Ma et al. [24] propose to employ Adaptive Instance Normalization [52] to transfer the style latent from the exemplar image. Park et al. [25] proposed a novel normalization layer for image synthesis and solved the problem of vanishment of semantic map of sparse input on synthesis in the previous image synthesis task. In contrast to the above approach of passing only global styles, our approach is to pass fine-grained local styles from the semantic counterpart region of the exemplar through the proposed self-attention mechanism.

Our work is inspired by recent examples-based image coloring, but we address a more subtle problem: exemplar-based coloring of sparse semantic and informationally complex sketches. At the same time, we present a novel training scheme to learn visual cross-domain correspondence and a sound adversarial strategy designed for sketch-based tasks aiming to improve the final imaging quality.

3. Proposed Method

In this section, we will describe the details of the proposed methods as shown in Figure 1. We first introduce a learnable domain alignment network in which dense semantic correspondences can be established, where the CSFT module is used to find spatial-level correspondences between the inputs. Then, we apply a coarse-to-fine generator to refine the coarse images gradually. Finally, we describe the structure and color strategy of the proposed discriminator.

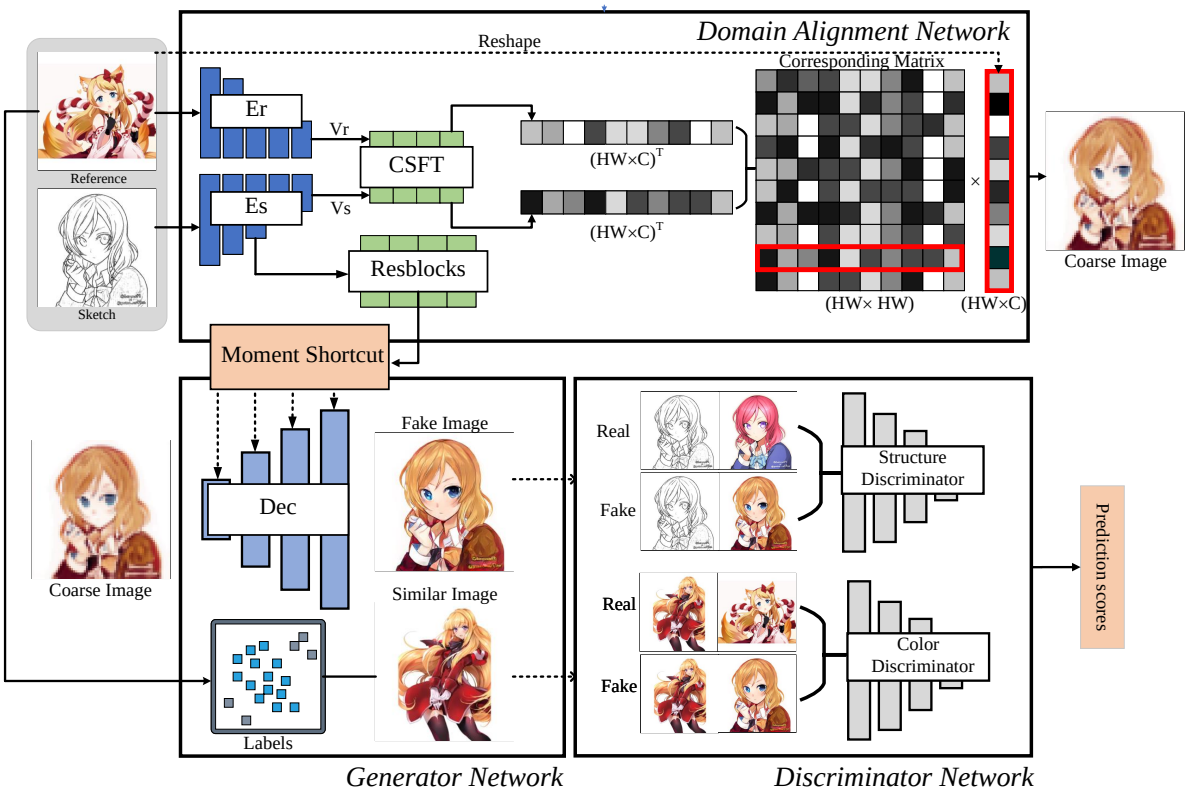

Figure 1. The illustration of the proposed framework. It contains three parts: Domain Alignment Network, Generator Network with Moment Shortcut strategy, and Discriminator Network with the structural and colorific conditions. Given the sketch input $x_s \in \mathbb{R}^{H\times W\times 1}$ and the exemplar input $y_e \in \mathbb{R}^{H\times W\times 3}$, the Domain Alignment Network adapts them into a common domain c, where the dense corresponding is established, to get the coarse outputs. Then, the generator refines the coarse images and outputs the refined images.

3.1. Domain Alignment Network

Image analogy [28,29,53] is a typical style migration method that uses a pre-trained VGG network to propose high-level abstract semantic information and find a suitable match on the target image (e.g., a realistic photo converted to a painting under the same semantic target). However, this approach does not apply to the migration task of sketches since sketches contain only a limited binary structure. The traditional VGG layer cannot extract suitable features for matching. Therefore, we propose a domain alignment network to establish correspondence between sketches and examples. However, conventional domain alignment is problematic in obtaining common domains in different semantics and different styles, so we propose a cross-domain spatially feature transfer (CSFT) module to help solve this problem.

3.1.1. Domain Alignment

To be specific, we let user inputs $x_s \in \mathbb{R}^{H\times W\times 1}$, and $y_e \in \mathbb{R}^{H\times W\times 3}$, s denote the domain of sketch, e denotes the domain of exemplar, and H, W denote the height and width, respectively. Additionally, we construct exemplar training pairs by using paired data $\{x_s, x_e\}$ that are semantically aligned but differ in domains. Similarly, exemplar training pairs $\{y_e, y_s\}$ are constructed in the same way as shown in Figure 2. Firstly, we project the given inputs x_s and y_e into a common domain c where the representation is able to represent the semantics for both distinct input domains. Let $\mathcal{F}(x_s), \mathcal{F}(y_e)$ be the corresponding features of x_s, y_e, where $\mathcal{F}(\cdot) \in \mathbb{R}^{H\times W\times L}$, L denotes the producing L activation maps $(f^1, f^2, \ldots, f^L)$, and H,W are feature spatial size. Then, we let $\mathcal{F}_{s\to c}$ and $\mathcal{F}_{e\to c}$ be representations of the feature embedding, where the embedding space is the common domain c. So, the presentation can be formulated as:

$$x_c = \mathcal{F}_{s\to c}(x_s; \theta_{\mathcal{F},s\to c}) \tag{1}$$

$$y_c = \mathcal{F}_{e\to c}(y_e; \theta_{\mathcal{F},e\to c}) \tag{2}$$

where θ denotes the learnable parameter of feature layers. The representations x_c and y_c contain the semantic and stylistic features of the inputs. In practice, domain alignment is crucial for correspondence establishment because x_c and y_c can be further matched with specific similarity measures in the same domain. Therefore, how to draw the representations of x_c and y_c more closely is a critical issue.

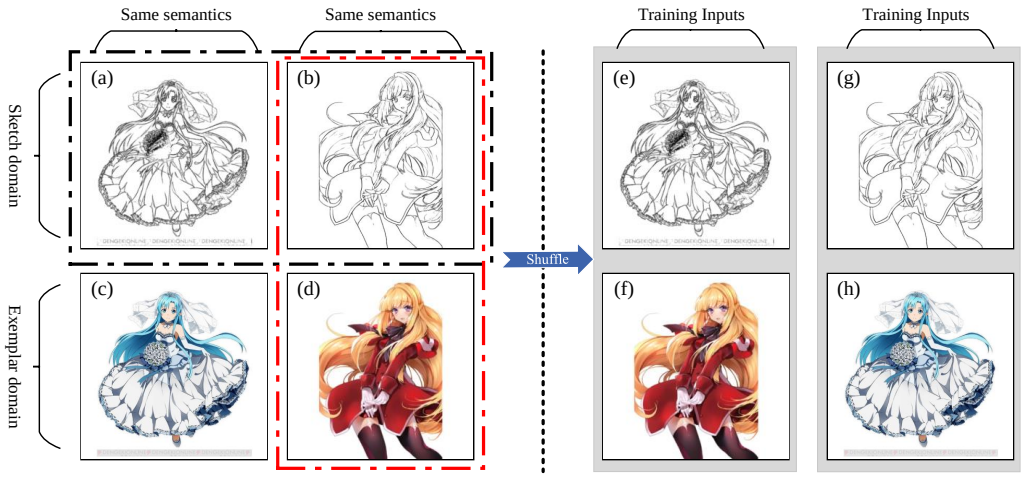

Figure 2. The illustration of training pairs. We construct pairs data $\{x_s, x_e\}$ (**a**,**c**), $\{y_e, y_s\}$ (**b**,**d**). In the training phase, we will shuffle the data as the training pair inputs (e.g., **e**–**h**). Subscript e means exemplar domain, and Subscript s means sketch domain.

3.1.2. Dense Correspondence

This subsection will describe how to close the distance between the features x_c, y_c obtained in the previous section. We use the cosine distance proposed by Zhang [27], which has the advantage of closing the intra-class distance and distancing the inter-class differences. Now, our goal is to build a learnable module to find the correlation matrix $\mathcal{M} \in \mathbb{R}^{HW \times HW}$, which can record the spatial correspondence between the representations. Let $i \in \{(H, W)\}$, $j\{(H, W)\}$ denote spatial positions of channel-wise centralized feature $\hat{x_c} \in \mathbb{R}^C$ and $\hat{y_c} \in \mathbb{R}^C$. Therefore, the formula can be written as:

$$\mathcal{M} = \frac{\hat{x}_c(i)^T \cdot \hat{y}_c(j)}{\|\hat{x}_c\| \cdot \|\hat{y}_c\|_2} \tag{3}$$

where $\hat{x}_c(i) = x_c(i) - mean(x_c(i))$ and $\hat{y}_c(j) = y_c(j) - mean(y_c(j))$. The matrix $\mathcal{M}$ indicates a dense pixel-by-pixel spatial correspondence.

To establish an efficient spatially dense correspondence, we also need an efficient feature transfer module intending to map different local features of the input to valid regions. We do not apply direct supervised learning to the domain alignment network, but indirect joint training through a proposed Dynamic Moment Shortcut method, which allows the entire architecture to preserve end-to-end optimization capabilities. In this way, the transformation network may find that high-quality coloring images can only be produced by correct domain mapping of the exemplar input, which explicitly compels the network to learn the accurate dense correspondence. In light of this, we let $w_{y \to x}$ by matching and computing the most relevant pixels in y_e and matrix $\mathcal{M}$ in the shared domain c.

$$w_{y \to x}(i) = \sum_{j}^{HW} softmax(\alpha \mathcal{M}(i, j) \cdot y'_e) \tag{4}$$

where α denotes a coefficient to control the degree of soft smoothing, default is 100. $y'_e \in \mathbb{R}^{HW}$ is the deformed vector of y_e.

3.1.3. Cross-Domain Spatially Feature Transfer

Under the guidance of Equation (4), we, therefore, propose the Cross-domain Spatially Feature Transfer module, which can effectively facilitate the establishment of spatially dense correspondence to the global statistical relationship between input features as shown in Figure 3.

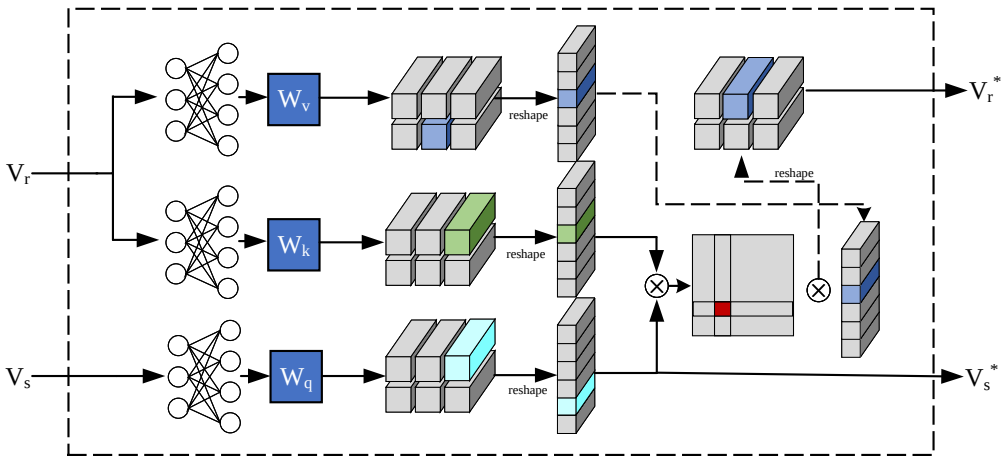

Figure 3. The illustration of the cross-domain spatially feature transfer (CSFT) module. CSFT establishes the dense correspondence mapping through the self-attention mechanism. The output results will be used for the next step of conversion, that is, to calculate the correspondence matrix $\mathcal{M}$.

To be begin with, each of the two feature pyramid networks Er and Es consists of L convolutional layers, producing L activation maps $(f^1, f^2, \ldots, f^L)$. Then, we apply

downsampling to each response layer f^i so that it scales to a consistent spatial size of f^L, and concatenate them along the channel dimensions, obtaining the organized activation feature map V, i.e.,

$$V = [\phi(f^1), \phi(f^2), \ldots, \phi(f^{L-1}), f^L] \tag{5}$$

where ϕ denotes the spatial downsampling method of each feature map, in this manner, we simultaneously obtained semantic information from high to low inputs.

Then, we reshape V in to $\hat{V} = [v^1, v^2, \ldots, v^{hw}] \in \mathbb{R}^{d_v \times HW}$, where $v^i \in \mathbb{R}^{d_v}$ means the spatial flatten representation of the i-th vector in V and $d_v = \sum_L^{l=1} channle(l)$. Then we can get v_s^i in $\hat{V}_s$ and v_r^j in $\hat{V}_r$, as indicated below:

$$\hat{V}_s = [v_s^1, v_s^2, \ldots, v_s^{hw}], v_s^i \in d_v \tag{6}$$

$$\hat{V}_r = [v_r^1, v_r^2, \ldots, v_r^{hw}], v_s^i \in d_v \tag{7}$$

After that, given the v_s^i and v_r^j, we can obtain the self-attention matrix $\mathcal{A} \in hw \times hw$, and following [54], we can get the scaled dot product result of α_{ij}:

$$\alpha_{ij} = softmax\left(\frac{W_q v_s^i \cdot W_k v_r^j}{\sqrt{d_v}}\right) \tag{8}$$

where $W_q, W_k \in \mathbb{R}^{d_v \times d_v}$ represents multilayer perceptron, and $\sqrt{d_v}$ denotes the scaling factor. α can be used as the calculated attention weight of how much information v_s^i should bring from v_r^j. Now, we can obtain the context vector V^* of region i of the exemplar image.

$$V^* = \sum_j \alpha_{ij} W_v v_r^j \in \mathbb{R}^{d_v \times hw} \tag{9}$$

Then, the dimension of V^* is adjusted by operations such as 1×1 convolution to obtain the x_c, y_c.

3.2. Coarse-to-Fine Generator

We employ a coarse-to-fine generative architecture to jointly train the domain alignment network, providing end-to-end training capability for the model. To avoid the failure of coarse image generation, we incorporate a Dynamic Moment Shortcut (DMS) structure in the generator, which has been shown to facilitate the generation of coarse deformation images.

Dynamic Moment Shortcut

Inspired by Dynamic Layer Normalization [55,56] and Position Normalization [33], and we employ Dynamic Moment Shortcut (DMS) in our generator. In generative models, although the conventional regularization layer may promote model convergence, it eliminates important semantic information about the images, which may cause generation failures, making it necessary for decoder structures with huge parameters to relearn the feature maps.

Instead, the introduction of DMS injects the positional moments extracted from earlier layers into the later layer of the network, enabling joint training of domain alignment networks with a low parametric number of decoders.

3.3. Structural and Colorific Strategy

In order to improve the color quality of the sketches, we propose the colorific and structural strategy, which effectively contributes to excellent and aesthetic coloring results. Here next, we describe in detail the structural and colorific strategy.

3.3.1. Structural Condition

The structural conditions are a brief overview and representation of the objects. We represent them using a series of binary black and white images, also the sketches we refer to. Concretely, we apply xDoG [57] in the training phase to generate simulation sketches, which can constitute our structural conditions.

We train the discriminator by composing the structural information of the exemplar and the generated samples, respectively, letting the discriminator focus on comparing the structural reasonableness of the generated images and maintaining consistency with the sketches. The ablation experiments show that the structural discriminator can reduce the occurrence of color diffusion.

3.3.2. Colorific Condition

The color condition indicates whether the image's color matches the example image, and it is the key to generating reasonable colors. Our model strives to generate a reasonable coloring result given a sketch image and a reference image. We apply the multi-scale discriminators in the discriminator network and use image processing techniques to extract sketches and color styles from these RGB images automatically.

In the following way, we compute a 3D lab color histogram ($8 \times 8 \times 8$) for each RGB image [30] and then measure their similarity by k-means clustering if their colors are close to each other to merge the exemplar images. As shown in Figure 4, we get an image with similar color similarity to the reference input as our color conditional input. In this way, the discriminator improves its sensitivity to the color correlation of the generated images and exemplar.

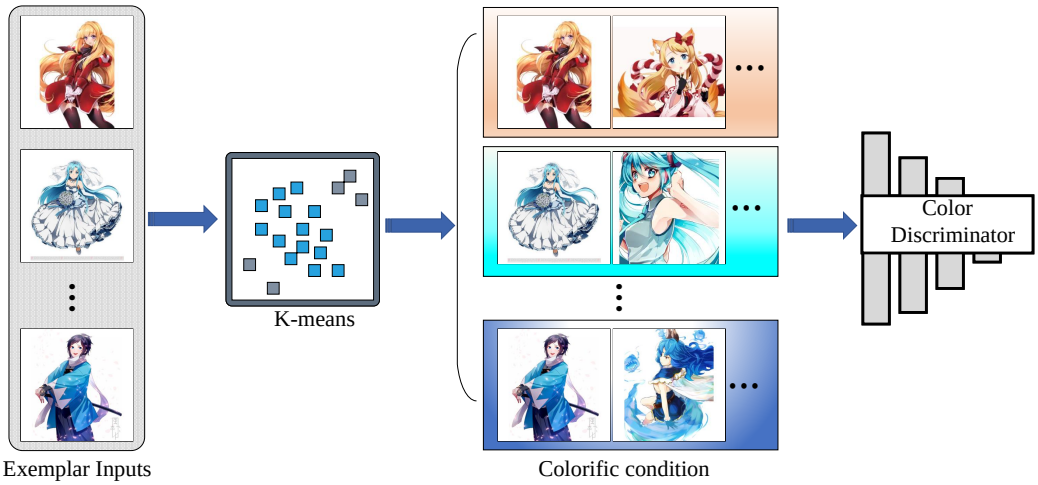

Figure 4. We cluster the images with similar hues by the K-means method. Then, we can obtain colorific conditions and use them in the discriminator in favor of K-means. Ablation experiments show that color conditions can effectively improve the quality of generated images.

3.3.3. Structural and Colorific Discriminators

As shown in Figure 5, we use pairwise discriminators with structural and colorific conditions to jointly train the generator part. Specifically, the structural discriminator is responsible for determining whether the generated images are structurally plausible and maintain structural consistency with the sketch input. We carefully designed positive and negative sample pairs to compel them to be sensitive only to the resulting structure. The colorific discriminator is responsible for identifying whether the resulting colors are reasonable. We perform positive and negative samples on images with different structures but similar colors, which forces the color discriminator to be more sensitive to changes in color patterns and promotes the generation of images that retain more of the style from the exemplar input. The structure discriminator prefers the spatial scale, while the coloring discriminator focuses on the style domain.

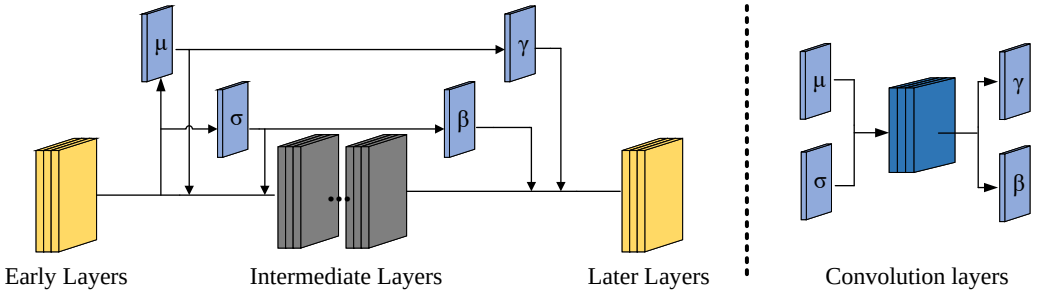

Figure 5. (**Left**) injecting the extracted mean and standard deviation as β and γ. (**Right**) one may employ learnable convolution layers to predict modulated β and γ dynamically based on μ and σ.

3.4. Loss for Exemplar-Based Sketch Colorization

We jointly train the domain alignment network and generator network along with the following loss functions.

3.4.1. Loss for Exemplar Translation

As shown in previous work [58], perceptual loss penalizes the model to decrease the semantic gap in the generated output, which means the multi-scale spatial differences of intermediate activation feature maps between the generated output and ground truth from the pre-trained VGG network.

$$\mathcal{L}_{perc} = \|\phi(G(x_s, y_e)) - \phi(x_e)\|_1 \tag{10}$$

where ϕ denotes the activation feature maps of l-th layer extracted at the $relu5\_2$ from the pre-trained VGG19 network.

Sajjadi et al. [59] have shown that reducing the style loss of the difference between the covariances of the activation maps helps to resolve the checkerboard effect. Therefore, we applied style loss to facilitate style transfer from the exemplars as follows:

$$\mathcal{L}_{style} = \mathbb{E}[\|\mathcal{G}(\phi(G(x_s, y_e))) - \mathcal{G}(\phi(y_e))\|_1] \tag{11}$$

where $\mathcal{G}$ denotes the gram matrix.

Meanwhile, we employ the contextual loss the same as [60] to let the output adopt the style from the semantically corresponding patches from y_e.

$$\mathcal{L}_{context} = \sum_l \omega_l \left[-log(\frac{1}{n_l} \sum_i \max_j A^l(\phi_i^l(G(x_s, y_e), \phi_j^l(y_e)))) \right] \tag{12}$$

where i and j indexes the feature map of layer ϕ^l, which contains the n_l feature maps and w_l restrains relative importance of different layers. In contrast to style loss, which primarily utilizes high-level features, context loss uses $relu2\_2$ through $relu5\_2$ layers because low-level features capture richer style information (e.g., color or texture) used to convey exemplar appearance.

3.4.2. Loss for Pseudo Reference Pairs

We construct training exemplar pairs $\{x_s, x_e\}$ that are semantically aligned but domain separated. Concretely, we apply random geometric distortion such as TPS transformation $s(\cdot)$, a non-linear spatial transformation operator to x_e, and get the distorted image $x'_e = s(x_e)$. This keeps our model from lazily bringing the color in the same spatial position from x_e. The interpretation of x_s should be its counterpart x_e when considering x_e as an exemplar. We proposed to penalize the pixel-wise difference between the output and the ground truth x_e as below:

$$\mathcal{L}_{pseudo} = \mathbb{E}[\|G(x_s, y_e) - x'_e\|_1] \tag{13}$$

3.4.3. Loss for Domain Alignment

We need to ensure that the representations x_c and y_c are in the same domain to make the domain alignment meaningful. To achieve this, we use the pseudo exemplar pair $\{x_s, x_e\}$ and $\{y_s, y_e\}$ to establish a shared domain c by penalizing the L1 distance between the representations.

$$\mathcal{L}_{align} = \|\mathcal{F}_{s\to c}(x_s) - \mathcal{F}_{e\to c}(x_e)\|_1 + \|\mathcal{F}_{s\to c}(y_s) - \mathcal{F}_{e\to c}(y_e)\|_1 \tag{14}$$

In this way, the model can gradually learn the mapping of different domains to a common domain.

3.4.4. Loss for Adversarial Network

We proposed to train a conditional discriminator [34] with the structural and colorific conditions to discriminate the translation output and the ground truth sample from distinct domains. We construct the discriminator input as in Section 3.3.

$$\begin{aligned}\mathcal{L}_{adv} = \mathbb{E}[logD_s(y_s, x_s) + logD_r(I_{similar}, x_e)] \\ +\mathbb{E}[log(1 - D_s(G(x_s, x_e), x_s)) \\ +log(1 - D_r(G(x_s, x_e), x_r))]\end{aligned} \tag{15}$$

where $I_{similar}$ denotes the sample that is similar to exemplar inputx_e in color.

4. Experiments

This section demonstrates the superiority of our approach on a range of domain datasets, including real photos and anime (comics).

4.1. Implementation

We implement our model with the size of input images fixed in 256 × 256 resolution on every dataset. For training, we adopt the Adam solver for optimization with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and the learning rates are both initially set to 0.0001 for generator and discriminator, respectively, following TTUR [61]. We conduct the experiments using NVIDIA GeForce RTX 3090 with batch size set as 8, and it probably takes three days to train 100 epochs on the Animepair dataset.

4.2. Dataset

4.2.1. Anime-Sketch-Colorization-Pair Dataset

We use Kaggle's anime-sketch-colorization-pair [62] dataset to train our model to validate the model's performance on hand-drawn data. It contains 14,224 training samples and 3545 test samples, including paired hand-crafted sketch images and corresponding color images.

4.2.2. Animal Face Dataset

The Animal Face Dataset [63] includes 16,130 high-quality animal face data containing several distinct domains of animal species, namely cats, dogs, and wild animals, with wild animals including lions, tigers, foxes, and other animals. We use this dataset to validate the model's performance in cross-domain image translation, and it turns out that our model can work well.

4.2.3. Edge2Shoe Dataset

Edge2Shot [64,65] contains paired sketch color shoe images that have been widely used for image-to-image conversion tasks. With this dataset, we can effectively evaluate the performance of our method and existing methods on unpaired image-to-image transformation tasks.

4.3. Comparisons to Baselines

We select different state-of-the-art image translation methods for visual comparison. (1) CycleGAN, a leading unsupervised image translation method. (2) MUINT, a multi-modal unsupervised image translation framework. (3) SPADE, an advanced framework for semantic image translation. (4) Sun et al.'s method, a recent reference-based sketch coloring method with good results on an icon dataset. (5) Cocos Net, an exemplar-based cross-domain image translation method for domain alignment using the learned shared embedding space.

4.4. Quantitative Evaluation

The quantitative model performance on different datasets is shown in Table 1. We evaluate our proposed method from five aspects:

- Firstly, we use Fréchet Inception Distance (FID) [61] to measure the distance between the synthetic image and the natural image distribution. FID calculated the Wasserstein-2 distance between the two Gaussian distributions in line with the features representation of a pre-trained convolution network InceptionV3 [66]. As Table 2 shows, compared with other excellent models, our proposed model has the best score in FID.
- Peak Signal to Noise Ratio (PSNR) is an engineering term representing the ratio between a signal's maximum power and the destructive noise power that affects its fidelity. We also evaluate the PSNR index of the models on different datasets, as shown in Table 3, and our model has achieved good performance.
- Structural Similarity (SSIM) [67] is also an image quality evaluation index, which measures the similarity of two images from three aspects: brightness, contrast, and structure. The larger the value, the better, and the maximum is 1. The quantitative results are shown in Table 4.
- NDB [68] and JSD [69]. To measure the similarity of the distribution between the real and generated images, we used two bin-based metrics, NDB (Number of Statistically-Different Bins) and JSD (Jensen-Shannon Divergence). These metrics evaluate the degree of pattern missing in the generated model. Our model has achieved good performance, as shown in Table 5.

Table 1. Model performance on FID, PSNR, SSIM, NDB, and JSD metrics. The arrow direction represents the better numerical direction of the metric (e.g., smaller FID, better performance).

	Dataset	**FID ↓**	**PSNR ↑**	**SSIM ↑**	**NDB ↓**	**JSD ↓**
Animal Faces	Cat	25.64	11.90	0.53	2.21	0.018
	Dog	26.65	12.77	0.62	2.54	0.021
	Wild	27.41	11.96	0.64	3.12	0.028
Comics	Anime-pair	19.14	16.44	0.83	2.00	0.016
Hand-drawn	Edge2shoe	15.69	16.72	0.83	2.01	0.015

Table 2. Model performance on metric of FID. sc means the structural condition, and cc means the colorific condition. Bold means best performance.

	Animal Face			Comics	Hand-Drawn
Methods	**Cat**	**Dog**	**Wild**	**Anime-Pair**	**edge2shoe**
SPADE	42.52	37.39	47.41	58.62	32.55
MUINT	33.48	32.45	42.54	37.45	29.47
CycleGAN	70.44	80.54	88.19	106.45	70.96
Sun et al.	48.45	45.45	55.69	67.65	38.46
Cocos Net	29.47	30.11	27.56	24.93	19.64
Ours(w/o *cc*)	28.12	26.42	29.13	28.13	18.77
Ours(w/o *sc*)	30.34	30.58	33.65	24.95	22.98
Ours(w/o CSFT)	33.54	36.21	34.21	30.96	24.16
Ours($full$)	**25.64**	**26.65**	**27.41**	**19.14**	**15.69**

Table 3. Model performance on metric of PSNR. Bold means best performance.

	Animal Face			Comics	Hand-Drawn
Methods	**Cat**	**Dog**	**Wild**	**Anime-Pair**	**edge2shoe**
SPADE	9.89	7.68	9.54	11.57	10.15
MUINT	10.32	10.45	9.59	12.96	12.11
CycleGAN	8.47	8.21	7.68	10.11	10.01
Sun et al.	9.36	10.45	10.42	12.41	13.34
Cocos Net	11.21	11.44	11.69	14.65	**16.73**
Ours	**11.90**	**12.77**	**11.96**	**16.44**	16.72

Table 4. Model performance on metric of SSIM. Bold means best performance.

	Animal Face			Comics	Hand-Drawn
Methods	**Cat**	**Dog**	**Wild**	**Anime-Pair**	**edge2shoe**
SPADE	0.42	0.44	0.42	0.40	0.40
MUINT	**0.62**	0.60	**0.66**	0.71	0.70
CycleGAN	0.51	0.51	0.52	0.50	0.50
Sun et al.	0.52	0.61	0.59	0.70	0.71
Cocos Net	0.53	**0.62**	0.63	0.81	0.82
Ours	0.53	**0.62**	0.64	**0.83**	**0.83**

Table 5. Model performance on metric of NDB and JSD. Bold means best performance.

	Animal Face						Comics		Hand-Drawn	
Methods	**Cat**		**Dog**		**Wild**		**Anime-Pair**		**edge2shoe**	
	NDB	**JSD**	**NDB**	**JSD**	**NDB**	**JSD**	**NDB**	**JSD**	**NDB**	**JSD**
SPADE	4.14	0.035	3.14	0.030	3.68	0.032	4.34	0.041	4.00	0.033
MUINT	2.25	0.020	3.01	0.029	**3.01**	0.029	3.51	0.029	2.54	0.019
CycleGAN	4.45	0.041	4.56	0.041	4.51	0.040	5.12	0.048	4.87	0.047
Sun et al.	4.41	0.040	4.11	0.039	3.47	0.035	3.41	0.030	3.28	0.020
Cocos Net	2.20	**0.018**	2.59	0.022	**3.01**	**0.024**	2.36	0.018	**2.01**	**0.015**
Ours	**2.21**	**0.018**	**2.54**	**0.021**	3.12	0.028	**2.00**	**0.016**	**2.01**	**0.015**

4.5. Qualitative Comparison

Figure 6 provides a qualitative comparisons of different approach. It shows that our proposed model exhibits the most visually appealing quality while preserving the style

of the examples better while retaining as much semantic information in the sketches as possible, compared to prior coloring approaches. This also correlates with the quantitative results, where we show the visual performance of our model under different datasets in Figures 7–9.

Figure 6. Qualitative results with existed colorization methods on anime datasets. All results are generated from the unseen dataset with sketch input and exemplar image under random selection within the validation set.

Figure 7. Qualitative results of our method on the edge2shoe dataset. Each row has the same semantic content, while each column has the same reference style.

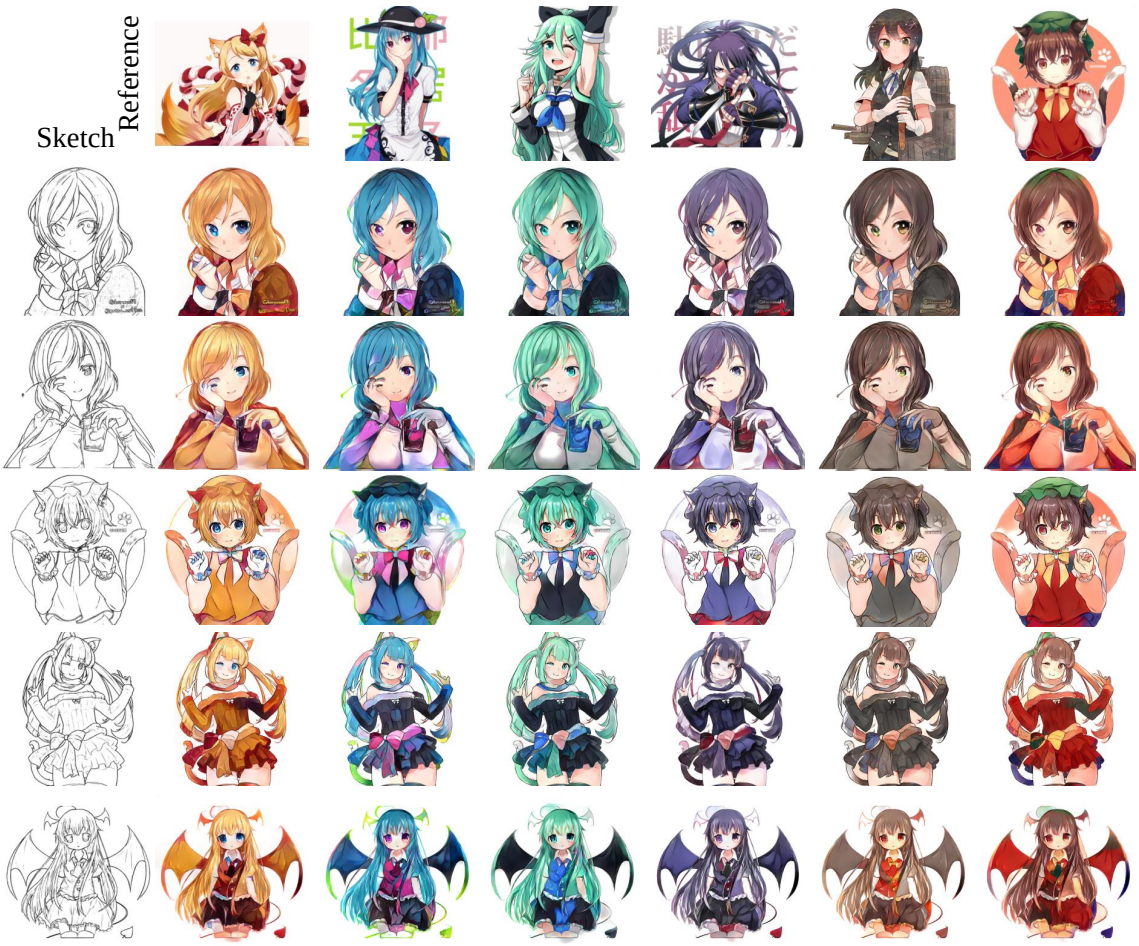

Figure 8. Qualitative results of our method on the anime dataset. Each row has the same semantic content, while each column has the same reference style. Please note that all the above results are generated from unseen images because the goal of our task is not to reconstruct the original image.

Figure 9. Qualitative results of our method on the animal-face dataset. Each row has the same semantic content, while each column has the same reference style.

4.6. Ablation Study

In order to verify the effectiveness of each part, we organized tailored ablation experiments. As Table 6 shows, domain alignment loss $\mathcal{L}_{align}$ plays a crucial role in cross-domain image translation, which not only effectively facilitates training, but also generates satisfying images. We also ablate the contextual loss $\mathcal{L}_{context}$. In our experiments, we found that although the network produced the final output, the feature correspondence may have a large mismatch, and using $\mathcal{L}_{context}$ loss enabled the correspondence to be well established.

Table 6. FID scores according to the ablation of loss function terms described in Section 3.4. Bold means best performance.

	Animal Face			Comics	Hand-Drawn
Loss Function	**Cat**	**Dog**	**Wild**	**Anime-Pair**	**edge2shoe**
w/o $\mathcal{L}_{context}$	40.68	52.65	50.52	33.51	32.69
w/o $\mathcal{L}_{pseudo}$	25.87	26.85	28.65	19.14	15.77
w/o $\mathcal{L}_{align}$	40.74	37.37	46.49	51.62	42.55
w/o $\mathcal{L}_{adv}$	42.51	38.39	47.44	58.68	44.55
full	**25.64**	**26.65**	**27.41**	**19.14**	**15.69**

As shown in Table 2, we performed ablation experiments for the proposed structural and colorific conditions. The experimental results prove that the strategy effectively reduces detail loss and color diffusion. As shown in Figure 10, the colorific condition can promote the correct matching of the exemplar style and sketch correspondence, and the structure condition can reduce mismatch and color diffusion.

Figure 10. A qualiative example presenting the effectiveness of structural and colorific conditions. (**a**) sketch input; (**b**) exemplar input; (**c**) output (w/o colorific condition); (**d**) coarse image (w/o colorific condition); (**e**) output (w/o structural condition); (**f**) coarse image (w/o structural condition) (**g**) output (full); and (**h**) coarse image (full).

As Table 2 shows, the FID metrics perform better with the addition of the CSFT module since CSFT can effectively facilitate the establishment of pixel-level correspondences and eliminate certain incorrect dense correspondences. At the same time, we found in practice that the addition of CSFT joint training can facilitate coarse image generation for domain-aligned networks. The control group of CSFT is a series of residual convolutions to maintain input–output invariance.

5. Discussion

The method proposed in this paper facilitates the solution of the problem of coloring sketches with sparse information. Traditional image translation or image transfer methods are not well suited for sketch colorization tasks because they have limited capability to establish correspondence between sparse semantic images and exemplars. Therefore, for sketch colorization, we propose a cross-domain alignment network that facilitates dense correspondence at the pixel scale using the proposed CSFT module, and the proposed structural and colorific conditions can be effectively applied to exemplar-based sketch colorization tasks.

Our model is mainly trained on cropped image data with restricted resolution (e.g., 256×256). We do not employ a multi-scale architecture like pix2pixHD [36] for high-resolution image synthesis. Moreover, the model is not exhaustive, and it is difficult to establish a perfect and correct correspondence because of the diversity and uncertainty of the user input. Therefore, it is challenging for the model to learn how to determine the given style's suitability and color it reasonably within a specific limit from the style of the user-given exemplar. For example, in the animal face dataset, we find that the converted results are not always satisfactory, which is firstly caused by the excessive differences between different species and secondly by the fact that the model is not yet able to establish a perfect dense correspondence, and in this case, how to generate aesthetically and intuitively appropriate results should be our consideration.

Currently, the model proposed in this paper has been initially tried in a sketch colorization task. We believe that the proposed model has good potential for cross-domain image translation tasks. In the future, we plan to extend the framework to the high-resolution domain and integrate style-consistent examples into the keyframes of video data.

6. Conclusions

In this paper, we present a cross-domain translation framework for exemplar-based sketch colorization tasks. We propose the cross-domain alignment module, effectively establishing correspondence between isolated domains. In order to further promote cross-domain learning, we propose a pixel-wise feature transfer component based on the self-attention mechanism, which is called the cross-domain spatially feature transfer module (CSFT). At the training stage, we design a simple and effective strategy to term the structural and colorific conditions, which can effectively promote image quality. Our method achieves better performance than existing methods in both qualitative and quantitative experiments. In addition, our method learns dense correspondences of sketch images, paving the way for some interesting future applications, which shows the significant potential in the practice of content creation and other fields.

Author Contributions: Conceptualization, J.C.; Methodology, J.C.; Writing—original draft, H.Z.; Software, H.Z.; Data curation, H.L.; Validation, Y.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Opening Project of Guangdong Province Key Laboratory of Computational Science at the Sun Yat-Sen. University. 2021011. This project was supported by Guangzhou Key Laboratory of Intelligent Agriculture (201902010081).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GAN	Generative Adversarial Networks
CSFT	Cross-domain Spatially Feature Transfer Module

References

1. Gatys, L.A.; Ecker, A.S.; Bethge, M. A neural algorithm of artistic style. *arXiv* **2015**, arXiv:1508.06576.
2. Gatys, L.; Ecker, A.S.; Bethge, M. Texture synthesis using convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*.
3. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2414–2423.
4. Wen, J.; Xu, Y.; Li, Z.; Ma, Z.; Xu, Y. Inter-class sparsity based discriminative least square regression. *Neural Netw.* **2018**, *102*, 36–47. [CrossRef] [PubMed]
5. Gatys, L.A.; Ecker, A.S.; Bethge, M.; Hertzmann, A.; Shechtman, E. Controlling perceptual factors in neural style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3985–3993.
6. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
7. Ci, Y.; Ma, X.; Wang, Z.; Li, H.; Luo, Z. User-guided deep anime line art colorization with conditional adversarial networks. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 22–26 October 2018; pp. 1536–1544.
8. Zhang, L.; Li, C.; Simo-Serra, E.; Ji, Y.; Wong, T.-T.; Liu, C. User-guided line art flat filling with split filling mechanism. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Event, 19–25 June 2021; pp. 9889–9898.
9. Hati, Y.; Jouet, G.; Rousseaux, F.; Duhart, C. Paintstorch: A user-guided anime line art colorization tool with double generator conditional adversarial network. In Proceedings of the European Conference on Visual Media Production, London, UK, 17–18 December 2019; pp. 1–10.
10. Yuan, M.; Simo-Serra, E. Line art colorization with concatenated spatial attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Event, 19–25 June 2021; pp. 3946–3950.
11. Zhang, R.; Zhu, J.-Y.; Isola, P.; Geng, X.; Lin, A.S.; Yu, T.; Efros, A.A. Real-time user-guided image colorization with learned deep priors. *arXiv* **2017**, arXiv:1705.02999.
12. Xiao, Y.; Zhou, P.; Zheng, Y.; Leung, C.-S. Interactive deep colorization using simultaneous global and local inputs. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 1887–1891.
13. Chen, J.; Shen, Y.; Gao, J.; Liu, J.; Liu, X. Language-based image editing with recurrent attentive models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8721–8729.
14. Kim, H.; Jhoo, H.Y.; Park, E.; Yoo, S. Tag2pix: Line art colorization using text tag with secat and changing loss. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9056–9065.
15. Zou, C.; Mo, H.; Gao, C.; Du, R.; Fu, H. Language-based colorization of scene sketches. *ACM Trans. Graph. (TOG)* **2019**, *38*, 1–16. [CrossRef]
16. Lee, J.; Kim, E.; Lee, Y.; Kim, D.; Chang, J.; Choo, J. 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, Seattle, WA, USA, 16–18 June 2020; pp. 5801–5810.
17. Bugeau, A.; Ta, V.-T.; Papadakis, N. Variational exemplar-based image colorization. *IEEE Trans. Image Process.* **2013**, *23*, 298–307. [CrossRef]
18. Charpiat, G.; Hofmann, M.; Schölkopf, B. Automatic image colorization via multimodal predictions. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 126–139.
19. Liu, X.; Wan, L.; Qu, Y.; Wong, T.-T.; Lin, S.; Leung, C.-S.; Heng, P.-A. Intrinsic colorization. In Proceedings of the ACM SIGGRAPH Asia 2008 Papers, Singapore, 10–13 December 2008; pp. 1–9.
20. Tomasi, C.; Manduchi, R. Bilateral filtering for gray and color images. In Proceedings of the Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), Bombay, India, 1 January 1998; pp. 839–846.
21. Winnemöller, H.; Olsen, S.C.; Gooch, B. Real-time video abstraction. *ACM Trans. Graph. (TOG)* **2006**, *25*, 1221–1226. [CrossRef]
22. Zhao, M.; Zhu, S.-C. Portrait painting using active templates. In Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering, Vancouver, BC, Canada, 5–7 August 2011; pp. 117–124.
23. He, M.; Chen, D.; Liao, J.; Sander, P.V.; Yuan, L. Deep exemplar-based colorization. *ACM Trans. Graph. (TOG)* **2018**, *37*, 1–16. [CrossRef]
24. Ma, L.; Jia, X.; Georgoulis, S.; Tuytelaars, T.; Gool, L.V. Exemplar guided unsupervised image-to-image translation with semantic consistency. *arXiv* **2018**, arXiv:1805.11145.
25. Park, T.; Liu, M.-Y.; Wang, T.-C.; Zhu, J.-Y. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2337–2346.

26. Qi, X.; Chen, Q.; Jia, J.; Koltun, V. Semi-parametric image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8808–8816.
27. Zhang, B.; He, M.; Liao, J.; Sander, P.V.; Yuan, L.; Bermak, A.; Chen, D. Deep exemplar-based video colorization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 8052–8061.
28. Hertzmann, A.; Jacobs, C.E.; Oliver, N.; Curless, B.; Salesin, D.H. Image analogies. In Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, 12–17 August 2001; pp. 327–340.
29. Liao, J.; Yao, Y.; Yuan, L.; Hua, G.; Kang, S.B. Visual attribute transfer through deep image analogy. *arXiv* **2017**, arXiv:1705.01088.
30. Sun, T.-H.; Lai, C.-H.; Wong, S.-K.; Wang, Y.-S. Adversarial colorization of icons based on contour and color conditions. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 683–691.
31. Wang, M.; Yang, G.-Y.; Li, R.; Liang, R.-Z.; Zhang, S.-H.; Hall, P.M.; Hu, S.-M. Example-guided style-consistent image synthesis from semantic labeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 1495–1504.
32. Wen, J.; Zhang, Z.; Xu, Y.; Zhang, B.; Fei, L.; Liu, H. Unified embedding alignment with missing views inferring for incomplete multi-view clustering. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, Hawail, USA, 27 January–1 February 2019; Volume 33, pp. 5393–5400.
33. Li, B.; Wu, F.; Weinberger, K.Q.; Belongie, S. Positional normalization. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
34. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
35. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
36. Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8798–8807.
37. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
38. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1857–1865.
39. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
40. Liu, M.-Y.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
41. Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8789–8797.
42. Yelamarthi, S.K.; Reddy, S.K.; Mishra, A.; Mittal, A. A zero-shot framework for sketch based image retrieval. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 300–317.
43. Dutta, A.; Akata, Z. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5089–5098.
44. Chen, W.; Hays, J. Sketchygan: Towards diverse and realistic sketch to image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9416–9425.
45. Lu, Y.; Wu, S.; Tai, Y.-W.; Tang, C.-K. Image generation from sketch constraint using contextual gan. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 205–220.
46. Liu, F.; Deng, X.; Lai, Y.-K.; Liu, Y.-J.; Ma, C.; Wang, H. Sketchgan: Joint sketch completion and recognition with generative adversarial network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5830–5839.
47. Frans, K. Outline colorization through tandem adversarial networks. *arXiv* **2017**, arXiv:1704.08834.
48. Zhang, L.; Ji, Y.; Lin, X.; Liu, C. Style transfer for anime sketches with enhanced residual u-net and auxiliary classifier gan. In Proceedings of the 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR), Nanjing, China, 26–29 November 2017; pp. 506–511.
49. Lian, J.; Cui, J. Anime style transfer with spatially-adaptive normalization. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
50. Liu, M.; Ding, Y.; Xia, M.; Liu, X.; Ding, E.; Zuo, W.; Wen, S. Stgan: A unified selective transfer network for arbitrary image attribute editing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
51. Huang, X.; Liu, M.-Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 172–189.
52. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.
53. Lee, J.; Kim, D.; Ponce, J.; Ham, B. Sfnet: Learning object-aware semantic correspondence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2278–2287.
54. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.

55. Chen, T.; Lucic, M.; Houlsby, N.; Gelly, S. On self modulation for generative adversarial networks. *arXiv* **2018**, arXiv:1810.01365.
56. Kim, T.; Song, I.; Bengio, Y. Dynamic layer normalization for adaptive neural acoustic modeling in speech recognition. *arXiv* **2017**, arXiv:1707.06065.
57. Winnemöller, H.; Kyprianidis, J.E.; Olsen, S.C. Xdog: An extended difference-of-gaussians compendium including advanced image stylization. *Comput. Graph.* **2012**, *36*, 740–753. [CrossRef]
58. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; Ebrahimi, M. Edgeconnect: Structure guided image inpainting using edge prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27 October–2 November 2019.
59. Sajjadi, M.S.; Scholkopf, B.; Hirsch, M. Enhancenet: Single image super-resolution through automated texture synthesis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4491–4500.
60. Zhang, P.; Zhang, B.; Chen, D.; Yuan, L.; Wen, F. Cross-domain correspondence learning for exemplar-based image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Sealttle, WA, USA, 16–18 June 2020; pp. 5143–5153.
61. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
62. Kim, T. Anime Sketch Colorization Pair. Available online: https://www.kaggle.com/datasets/ktaebum/anime-sketch-colorization-pair (accessed on 1 June 2019).
63. Choi, Y.; Uh, Y.; Yoo, J.; Ha, J.-W. Stargan v2: Diverse image synthesis for multiple domains. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Sealttle, WA, USA, 16–18 June 2020.
64. Yu, A.; Grauman, K. Fine-Grained Visual Comparisons with Local Learning. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014.
65. Xie, S.; Tu, Z. Holistically-nested edge detection. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015.
66. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826.
67. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
68. Mao, Q.; Lee, H.-Y.; Tseng, H.-Y.; Ma, S.; Yang, M.-H. Mode seeking generative adversarial networks for diverse image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 1429–1437.
69. Fuglede, B.; Topsoe, F. Jensen-shannon divergence and hilbert space embedding. In Proceedings of the IEEE International Symposium onInformation Theory, ISIT 2004. Proceedings, Chicago, IL, USA, 27 June–2 July 2004; p. 31.