Region Assisted Sketch Colorization

Ning Wang, Muyao Niu, Zhihui Wang, *Member, IEEE*, Kun Hu, Bin Liu, *Member, IEEE*, Zhiyong Wang, *Member, IEEE*, and Haojie Li, *Member, IEEE*

***Abstract*— Automatic sketch colorization is a challenging task that aims to generate a color image from a sketch, primarily due to its inherently ill-posed nature. While many approaches have shown promising results, two significant challenges remain: limited color patterns and a wide range of artifacts such as color bleeding and semantic inconsistencies among relevant regions. These issues stem from the operation of traditional convolutional structures, which capture structural features in a pixel-wise manner, resulting in inadequate utilization of regional information within the sketch. Therefore, we propose the Region-Assisted Sketch Coloring (RASC) method, which introduces an intermediate representation called the 'Region Map' to explicitly characterize the regional information of the sketch. This Region Map is derived from the input sketch and is effectively formulated by our RASC architecture, enhancing the perception of region-wise features beyond the original pixel-wise features. Specifically, we start by employing the sketch encoder to extract hierarchical feature maps from the input sketches. Subsequently, we introduce a coarse-to-fine decoder comprising a series of Region-based Modulation (RM) blocks. This decoder modulates features that combine the modulation results of its previous block and the sketch features of the corresponding encoder block with our Region Formulation module. Each module explicitly formulates the sketch features in a region-wise manner. This accurately captures both the inner-region local style and inter-region global context dependency, resulting in various color patterns and fewer synthesis artifacts. Our experimental results show that our proposed method surpasses state-of-the-art methods in both synthetic and real sketch datasets.**

***Index Terms*— Sketch colorization, GAN, media art.**

I. Introduction

ANIME sketch colorization aims to generate a high-quality color image from a given sketch with sparse content. It has a wide range of applications in fields such as the animation industry, business advertising, and artistic design [1], [2]. However, it is very challenging due to its ill-posed nature, which means there exist multiple plausible results generated for one input. Due to the significant advancements and successes of deep learning techniques in various visual tasks, a wide range of deep learning methods have been thoroughly investigated in the field of colorization [3], [4], [5], [6], [7], [8], [9], [10], [11].

Current approaches in sketch colorization predominantly fall into two categories: reference-based methods [1], [3], [12], [13], [14] and non-reference based methods [5], [9], [15]. Reference-based methods require additional references such as color scribbles [3], [13], color images [12], [16], or text tags [14], [17], [18], [19] for line art colorization. These methodologies achieve colorization by discerning and transferring color information based on the inherent correlation between the input sketch and the provided color references. Recent advancements in non-reference-based sketch colorization have primarily leveraged Generative Adversarial Networks (GANs) [20]. Unlike reference-based methods, these techniques operate without relying on specific color guides for each sketch, which severely suffers from the ill-posed nature due to the absence of constraints in learning coherent color distribution mapping. Multi-model frameworks [15] have been investigated to generate diverse color styles, leveraging three distinctive models, each embodying specific color patterns. The concept of dual color space supervisions, exemplified by the incorporation of an additional HSV color space [5], has been embraced to harness the complementary information inherent in diverse color spaces. Additionally, the constraint of cycle-consistency [21] has been integrated into the colorization workflow, aiming to yield more natural and photorealistic illustrations [9]. CWR [14] introduces new constraints by contacting and discriminating skeleton region maps in a pixel-wise manner for sketch colorization. Significant progress has been made in the field of non-reference-based sketch colorization; however, the current methodologies are plagued by two primary limitations. Firstly, the existing methods tend to yield images characterized by a restricted variety of color patterns. Secondly, these methods underutilize the inherent structural information in sketches, resulting in the emergence of artifacts.

The essence of the observed limitation is ascribed to the operational characteristics of conventional convolutional structures, which capture structural information predominantly in a *pixel-wise* manner. Considering the inherent distinct and explicit regions in anime images, each sketch region typically presents an individual color style. However, conventional convolutional structures face significant challenges in learning a consistent style across all pixels within the same region and in learning the correlations in a *region-wise* manner. This capability is especially crucial for sketch colorization. Consequently, numerous artifacts are visible in the images generated, manifested as color bleeding (e.g., color gradients, wherein the color from one region permeates into adjacent regions) and semantic color inconsistency (e.g., the occurrence

Manuscript received 5 July 2023; revised 26 September 2023; accepted 15 October 2023. Date of publication 31 October 2023; date of current version 8 November 2023. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61976083, Grant 61932020, Grant 61772108, and Grant 1908210. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Aykut Erdem. *(Ning Wang and Muyao Niu contributed equally to this work.) (Corresponding author: Zhihui Wang.)*

Ning Wang is with the School of Software Technology, Dalian University of Technology, Dalian 116000, China.

Muyao Niu is with the Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo City, Tokyo 113-8654, Japan.

Zhihui Wang, Bin Liu, and Haojie Li are with the International School of Information Science and Engineering, Dalian University of Technology, Dalian 116000, China (e-mail: zhwang@dlut.edu.cn).

Kun Hu and Zhiyong Wang are with the School of Computer Science, The University of Sydney, Sydney, NSW 2008, Australia.

Digital Object Identifier 10.1109/TIP.2023.3326682

Fig. 1. Representative examples colorized by our RASC. The first row is the sketch input, and the second row is the corresponding colorization results. The third row is the demonstration for multiple colorization for the same input.

of disparate colors in two eyes or markedly distinct color styles within different segments of hair), as depicted in Figure 5.

Therefore, in this paper, we proposed the Region Assisted Sketch Colorization (RASC) method in pursuit of increasing the diversity of color patterns and reducing artifacts in the generated color images. Our RASC explicitly processes the sketch feature in a region-wise manner through region aggregation and region attention to effectively estimate the local style of each region and their implicit global dependencies. Specifically, our RASC includes a sketch encoder and a decoder which consists of several Region-based Modulation (RM) blocks. Initially, the structural information, represented as 'region maps,' is derived from the input sketch. Subsequently, the sketch is subjected to the encoder to generate a sequence of feature maps that are utilized by the decoder. This decoder, containing several RM blocks, integrates the modulated features from its preceding block and the sketch features corresponding encoder block. Importantly, this procedure is modulated by the corresponding region maps via the Region Formulation module, which formulates both the regional patterns and inter-region correlations to accurately depict local styles and global dependencies, mitigating colorization artifacts. Particularly, in the Region Formulation module, the hierarchical regional features are first obtained by aggregating the feature maps from each encoder block. Then, self-attentions are employed to formulate the correlations between the regions, ensuring a consistent style in a region. The obtained region-aware patterns are finally broadcast with the region map at the corresponding hierarchical level. Consequently, RASC can accurately estimate the local style of each region along with its implicit global dependencies, yielding visually convincing and color-rich images with fewer artifacts. In addition, the proposed architecture enables style-content disentanglement and multiple colorization results can be obtained. In summary, the key contributions of our proposed method are as follows:

- We propose an innovative Region Assisted Sketch Colorization (RASC) model designed for non-reference sketch colorization, with the objective of generating colorized images characterized by diversified and realistic color patterns. This model incorporates an additional region map, leveraged by the specifically devised region-aware architecture to optimize the colorization process.
- We devised a novel hierarchical Region-based Modulation (RM) Block that effectively formulates the regional patterns and their implicit correlations through the Region Formulation module, and introduces region-based modulation to modulate high-dimensional feature maps. Different from previous works, it explicitly capture structural information in a region-wise manner and accurately characterizes both the local style and the global context of regional patterns, alleviating synthesis artifacts in colorization results.
- Extensive experiments carried out on both synthetic and real sketch datasets validate that our proposed methodology surpasses existing state-of-the-art techniques. Further, the experiments showcase the potential for attaining controllable and seamless multi-colorization for a single sketch by manipulating the latent code.

II. Related Work

Non-reference-based sketch colorization is notably challenging due to the absence of explicit color references for sketches, which presents an ill-posed problem with minimal constraints for learning color distribution mappings. The initial attempt to address this challenge was made by Petalica Paint [15], which introduced three models: Tanpopo, Satsuki, and Canna. Each of these models utilized different U-Net [22]-based architectures to generate separated color styles for anime sketches. However, these models often exhibited limited diversity and perceptible synthesis artifacts in their colorization results. Ci et al. [4] introduced a local feature network aimed at mitigating overfitting to specific sketch types by extracting common semantic features from sketches of diverse styles. During the training phase, they also incorporated sketches both with and without color references to facilitate the learning of mappings applicable to both non-reference and reference-based colorization. Based on [4], HSV and RGB color spaces were integrated to construct dual color spaces for supervision [5]. DP loss and DCSA loss were introduced to guide the colorization process at both local and global dimensions respectively. Zhang et al. [9] incorporated the Cycle-Consistent idea from CycleGAN [21] to make network training more robust by applying two-way supervisions between sketches and color images. While considerable advancements have been made, significant gaps remain between generated colorizations and real anime, especially the inadequacy of diversity in color styles. Also, synthesis artifacts such as color bleeding and semantic color inconsistency continue to be prevalent issues in the generated results. CWR [14] integrates skeleton region maps sourced from DanbooRegion [23]. This integration serves to mitigate artifacts to some extent, owing to the implicit inclusion of region information within these maps. Differing from our method, we directly extract region maps from line art images. These extracted region maps provide explicit structural information about different regions in the line art, achieved by the Trapped Ball method [24] that performs fine-grained region partitioning. Notably, all of the aforementioned methods capture structural features in a pixel-wise manner, which constrains the network's capacity to fully comprehend the structural information of the sketches.

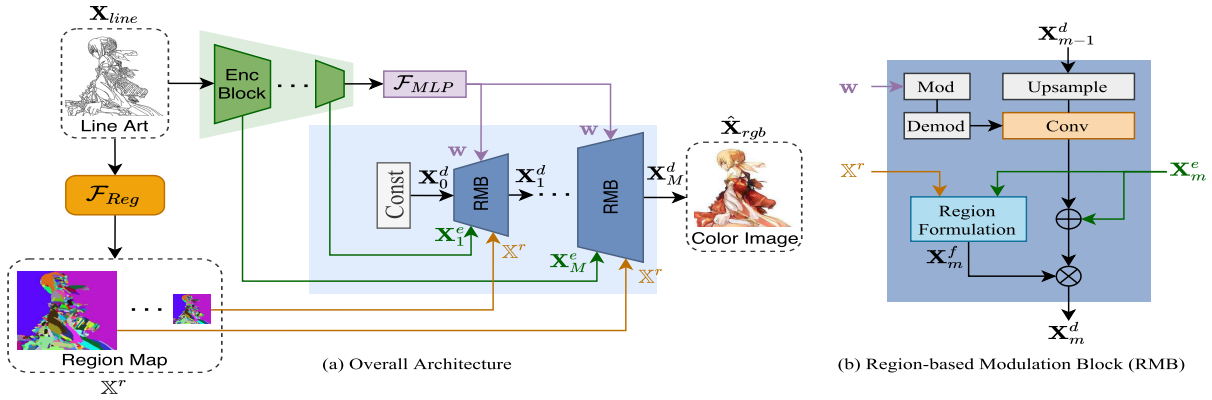


Fig. 2. Overview of the proposed network, which follows an encoder-decoder scheme. Given a sketch input, a latent style code, and a series of feature maps organized hierarchically based on the encoder blocks, they are subsequently fed into the region-based modulation (RM) Blocks within the decoder. Specifically, each RM Block initially employs the region formulation module to formulate the regional patterns and their correlations, utilizing the corresponding region map obtained directly from line art images. Subsequently, the block modulates the concatenated feature map accordingly. The intricate details pertaining to the region formulation module are depicted in Figure 3.

In contrast, our approach, which leverages the region map and our specialized region formulation module, empowered by region feature aggregation, region correlation, and region feature broadcast, facilitates a comprehensive comprehension of global contexts while preserving local patterns. Consequently, this empowers our model to produce colorization results that manifest superior coherency and diversity.

Reference-based sketch colorization methods often require additional references to alleviate the issues caused by the ill-posed nature of the colorization task. They usually exploit implicit feature correspondence between the sketches and the references such as color scribbles [3], [4], [5], [13], reference images [6], [7], [12], [16], and text tags [8], [17]. For example, Zhang et al. [3] proposed a two-stage colorization network to perform colorization as two simple sub-tasks with separate objectives, including drafting and refinement stages for improved colorization quality. Lian et al. [6] introduced Spatially-Adaptive Normalization (SPADE) to transfer semantic attributes from reference images to target images. Maejima et al. [7] proposed a graph-matching-based anime colorization method to colorize sketches using multiple reference images. Tag2Pix [8] was proposed to colorize a sketch according to the given color tags by devising a squeeze and excitation with a concatenation module to enhance multi-label segmentation for various semantic tags. The SGA method [25] enhances coloring quality by resolving gradient conflicts between the reference and target domains. ControlNet [18] and T2I-Adapter [19] leverage the capabilities of large-scale models for generating text to images, achieving high-quality colorization results with reduced network resource requirements. Nevertheless, these models are dependent on text references. When default text prompts are used or not available, the generated images tend to display a consistent color scheme.

Gray-scale image colorization is tasked with converting a gray-scale image into RGB space. Unlike sketches, gray-scale images possess dense structural information, facilitating networks in the formulation of spatial correlations and the recognition of patterns. For reference-based colorization, statistic similarities between the reference and the target were estimated through low-level similarities [26], [27], semantic features [28], or super-pixels [29], [30]. For automatic colorization [31], [32], [33], [34], [35], Su et al. [34] integrated instance-aware features to accommodate the appearance variations inherent to distinct objects. Wu et al. [35] recovered vivid color by leveraging the rich and diverse color priors encapsulated in the pre-trained BigGAN [36]. They extracted corresponding features utilizing a GAN encoder, subsequently integrating these features into the colorization process through feature modulations. Deoldify[1] was developed to automatically colorize world-realistic gray-scale images with several improvements such as self-attention and new training strategies. While these methods achieve notable colorization outcomes for gray-scale images, our empirical investigations reveal their inadequacy in appropriately colorizing sketches, due to the sparse nature of sketches and their lack of consideration for structural information in the region map.

III. Proposed Method

In this section, a detailed exposition of our RASC architecture is provided. Initially, a comprehensive overview of the model's entire workflow is presented. Subsequently, elaborations are made on its three pivotal components: (1) Region Map Identification, (2) Region-based Modulation (RM) Block, and (3) Region Formulation module. Finally, our loss functions are discussed in detail.

A. Overview

As illustrated in Figure 2(a), the proposed colorization architecture generates a colorized RGB image $\hat{\mathbf{X}}_{rgb} \in \mathbb{R}^{C_{rgb} \times W \times H}$ as the estimation of the ground truth $\mathbf{X}_{rgb}$ using a sketch input image $\mathbf{X}_{line} \in \mathbb{R}^{C_{line} \times W \times H}$, where $C_{rgb} = 3$ indicates the three RGB channels, $C_{line} = 1$ indicates the one-channel sketch, W is the width and H is the height of the images. The architecture first obtains the corresponding region maps $\mathbb{X}^r$, then introduces a CNN based encoder Enc to generate multi-level feature maps $\mathbb{X}^e = \{\mathbf{X}^e_M, \ldots, \mathbf{X}^e_1\}$ from the input sketch, where $\mathbf{X}^e_m \in \mathbb{R}^{c^e_m \times W^e_m \times H^e_m}$ is the output feature map

[1] https://github.com/jantic/DeOldify

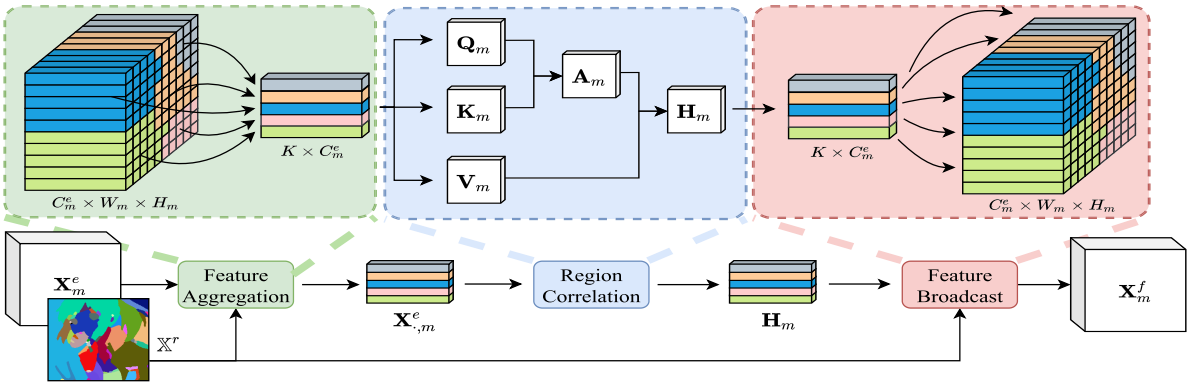


Fig. 3. Illustration of the proposed region formulation module. The module consists of three components: 1) region feature aggregation, 2) region correlation, and 3) region feature broadcast.

of the m-th encoder block and M is the number of encoder blocks in Enc, and a sketch related latent code $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ using an additional multilayer perceptron (MLP) on $\mathbf{X}_1^e$. Then, a coarse-to-fine decoder Dec involving $\mathbb{X}^e$, $\mathbf{w}$, and the region maps $\mathbb{X}^r$ is devised for the generation of the colorized image, which consists of several RM Blocks. Particularly, denote the output feature maps of Dec as $\mathbb{X}^d = \{\mathbf{X}_1^d, \ldots, \mathbf{X}_N^d\}$, where $\mathbf{X}_n^d \in \mathbb{R}^{C_n^d \times W_n^d \times H_n^d}$ is the output map of the n-th RM Block (see Section III-B.3) and N is the number of encoder blocks in Enc. We set $M = N$ in this study, i.e., the numbers of encoder blocks and RM Blocks are equal. Overall, $\mathbb{X}^e$ and $\mathbf{w}$ are in pursuit of two different perspectives. $\mathbf{w}$ serves as the style input of the decoder Dec and is mainly responsible for controlling the global style of the generated image (i.e., overall color style). While each $\mathbf{X}_k^e \in \mathbb{X}^e$ is taken by our RM Block as the multi-level injection mechanism of high-level representations to affect the local stochastic variations and guide the details of the structure reconstruction.

B. Three Region Aware Components

(1) Region Map Identification

In practical scenarios, photorealistic grayscale images typically contain abundant details, where explicit region divisions are not discernible. However, due to the unique nature of anime and the sparsity of the sketch, the line strokes partition the entire sketch into several distinct regions, each typically displaying its consistent color style. Therefore, we identify a series of region maps, denoted as $\mathbb{X}^r = \mathbf{X}_1^r, \ldots, \mathbf{X}_K^r$, to help define the regional patterns used to partition the input image into multiple regions. In detail, $\mathbf{X}_k^r \in \mathbb{R}^{W \times H}$ indicates the k-th region in $\mathbf{X}_{line}$ and K is the number of regions as illustrated in Figure 3. Note that if the pixel (i, j) of the input image is located in the k-th region, the element of the i-th row and j-th column in $\mathbf{X}_k^r$ satisfies $\mathbf{X}_k^r(i, j) = 1$; otherwise $\mathbf{X}_k^r(i, j) = 0$. These region maps can be obtained by using the trapped-ball algorithm [24]. Note that the scale of $\mathbf{X}_k^r$ may need to be resized to match the shape of $\mathbf{X}_m^e \in \mathbb{R}^{C_m^e \times W_m \times H_m}$ using nearest neighbour interpolation. For the sake of convenience, we do not introduce a new symbol for the resized $\mathbf{X}_k^r$ and $\mathbb{X}^r$.

(2) Region-based Modulation Block

The color images produced by existing conventional convolutional structures in a pixel-wise manner exhibit synthesis artifacts, leading to perceptually unsatisfying colorization results. This is attributed to color bleeding effects and semantic color inconsistency across corresponding regions. In order to address the aforementioned challenges, we introduce a novel mechanism termed 'Region-Based Modulation'. This mechanism incorporates individual region maps to guide the colorization process, consequently resulting in a significant reduction in synthetic artifacts. As depicted in Figure 2(b), the m-th Region Modulation (RM) Block is designed to receive four distinct inputs. These include the high-dimensional generative features, denoted as $\mathbf{X}_{m-1}^d$; the latent code, represented by $\mathbf{w}$; the region maps, indicated by $\mathbb{X}^r$; and finally, the high-dimensional features, $\mathbf{X}_m^e$, derived from the corresponding encoder block.

First, $\mathbf{X}_{m-1}^d$ is upsampled and put through the convolution block. Inspired by [37], we utilize the latent code $\mathbf{w}$ to modify the convolution block, which introduces global style information encapsulated in $\mathbf{w}$ to our RM Block. Then we use $\mathbf{X}_m^e$ and $\mathbb{X}^r$ to formulate region-aware patterns $\mathbf{X}_m^f$ through the Region Formulation module (see Section III-B). Finally, we use $\mathbf{X}_m^f$ to modulate the concatenation of $\mathbf{X}_m^e$ and the convolution output as:

$$\mathbf{X}_m^d = \mathbf{X}_m^f \odot (\mathbf{X}_m^e + \mathrm{Conv}(\mathrm{Upsample}(\mathbf{X}_{m-1}^d))). \quad (1)$$

It's important to note that prevailing methodologies, primarily relying on conventional convolutional neural structures, typically capture the structural information of sketches in a *pixel-wise* manner. Due to the inherent explicit regions of the sketch, achieving uniformity and consistency in both global and local patterns poses a significant challenge. Consequently, such methods are inefficient; They fail to accurately characterize the local style of a specified region, leading to undesirable color bleeding effects. Additionally, they fail to discern the global style correlations among disparate regions, resulting in semantic color inconsistencies. In contrast, our RM Block explicitly employs region assistance by formulating regional patterns and their correlations through the Region Formulation module. It then formulates high-dimensional sketch feature maps in a region-wise manner. This key difference provides a more robust and accurate colorization procedure and reduces artifact issues due to the unified correlation learning for all pixels located in the same region.

(3) Region Formulation

As illustrated in Figure 3, when provided with the region maps $\mathbb{X}^r$ and the high-dimensional features $\mathbf{X}^e_m$ from the corresponding encoder block, our Region Formulation module accomplishes the formulation of both the local style of each region and the implicit global correlations among them through feature aggregation and the incorporation of the self-attention mechanism.

1) Region Feature Aggregation: In order to formulate region-wise features, we need to aggregate representations located in the same region according to the region maps. In detail, the region maps $\mathbf{X}^r_k$ is adopted as a mask on $\mathbf{X}^e_m$ to obtain the features of the k-th region. Then, a maximum pooling strategy is applied to the spatial dimensions:

$$\mathbf{X}^e_{k,m}(c) = \max_{i,j}((\mathbf{X}^r_k \odot \mathbf{X}^e_m)(c,i,j)) \in \mathbb{R}^{C^e_m}, \tag{2}$$

where $\odot$ is an element-wise matrix multiplication for each channel of $\mathbf{X}^e_m$, $(\mathbf{X}^r_k \odot \mathbf{X}^e_m)(c,i,j)$ indicates the elements of the c-th channel, i-th row and j-th column in the feature maps, and $\mathbf{x}^e_{k,m}(c)$ is the c-th element in $\mathbf{x}^e_{k,m}$. Finally, by stacking $\mathbf{x}^e_{k,m}$ from $k=1$ to K in a row-wise manner, a feature matrix $\mathbf{X}^e_{:,m} \in \mathbb{R}^{K \times C^e_m}$ can be obtained as shown in Figure 3.

2) Region Correlation: Once region-wise features are acquired, we proceed to implement self attentions [38] on $\mathbf{X}^e_{:,m}$ to formulate patterns and establish correlations between these regions. In detail, query $\mathbf{Q}_m$, key $\mathbf{K}_m$ and value $\mathbf{V}_m$ matrices are formulated by linear projections:

$$\mathbf{Q}_m = \mathbf{X}^e_{:,m}\mathbf{W}_{Q,m}; \mathbf{K}_m = \mathbf{X}^e_{:,m}\mathbf{W}_{K,m}; \mathbf{V}_m = \mathbf{X}^e_{:,m}\mathbf{W}_{V,m}; \tag{3}$$

where $\mathbf{W}_{Q,m}$, $\mathbf{W}_{K,m}$, and $\mathbf{W}_{V,m}$ are matrices containing learnable parameters. Subsequently, attention weights are computed based on $\mathbf{Q}_m$ and $\mathbf{K}_m$, and an output region representation can be derived as follows:

$$\mathbf{H}_m = \text{softmax}(\frac{\mathbf{Q}_m\mathbf{K}^\intercal_m}{\sqrt{C^e_m}})\mathbf{V}_m, \tag{4}$$

where $\text{softmax}(\frac{\mathbf{Q}_m\mathbf{K}^\intercal_m}{\sqrt{C^e_m}})$ is denoted as $\mathbf{A}_m$ in Figure 3, representing the attention weight, and $\mathbf{H}_m \in \mathbb{R}^{K \times C^e_m}$ is derived as the final region patterns involving their correlations.

3) Region Feature Broadcast: Since each row of $\mathbf{H}_m$ contains attention-aware information for one region, we further broadcast each row to every pixel in the original feature map. To be more specific, the k-th row of $\mathbf{H}_m$ is broadcasted to populate a feature map $\mathbf{X}^f_m$ as follows:

$$\mathbf{X}^f_m(c,i,j) = \sum_{k=1}^{K} \mathbf{X}^r_k(i,j)\mathbf{H}_m(k,c), \tag{5}$$

where $c = 1,\ldots,C^e_m$, $i = 1,\ldots,W$ and $j = 1,\ldots,H$. Note that all within each region share identical regional patterns, which ensures consistency and alignment in terms of local style for each region throughout the modulation process, as detailed in Section III-B(2).

Fig. 4. Illustrations of the original image, synthetic sketches, mirror padding, and sketch simplification (from left to right).

C. Losses for Optimization

Overall, our RASC is defined as a function $\mathcal{G}$ with learnable parameters. For the optimization of $\mathcal{G}$, we employ both an adversarial loss and a perceptual loss [39], taking into account two distinct perspectives. The adversarial loss $\mathcal{L}_a$ plays a min-max game to guide the generation to follow the real color image distribution:

$$\min_{\mathcal{G}}\max_{\mathcal{D}} \mathcal{L}_a = \log(1+\exp(\mathcal{D}(\mathbf{X}_{rgb}))) + \log(1+\exp(-\mathcal{D}(\mathcal{G}(\mathbf{X}_{line})))), \tag{6}$$

where $\mathcal{D}$ is a convolution-based discriminator to classify whether a color image is generated or not.

The objective of the perceptual loss $\mathcal{L}_p$ is to maintain the overall structure of an input sketch and ensure the perceptual plausibility of the colorized image:

$$\min \mathcal{L}_p = \sum_{i=1}^{I} \left\| \phi_i\left(\mathcal{G}(\mathbf{X}_{line})\right) - \phi_i\left(\mathbf{X}_{rgb}\right) \right\|_1, \tag{7}$$

where ϕ_i denotes the activation map at the i-th layer of a pre-trained VGG-19 network [40]. Particularly, we chose the layers including $relu_{1-1}$, $relu_{2-1}$, $relu_{3-1}$, $relu_{4-1}$, and $relu_{5-1}$ from the VGG-19 network.

IV. Experimental Results and Discussions

A. Dataset

To train our model, we assembled a dataset comprising 40,000 triplets, each consisting of a sketch, a color illustration, and a region map. The color illustrations are collected from Danbooru2020 [41]. The sketches are extracted from color illustrations using XDoG algorithm [42] and the region maps are prepossessed using the algorithm mentioned in Section III-B(1). We set the parameters of XDoG as $\tau = 0.999997$, $\kappa = 1.0001$, and σ is randomly picked from $\{0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. Following the training phase, we employ three distinct test sets to comprehensively evaluate our proposed approach and existing methods. These test sets encompass both synthetic sketches and real sketches. (Test Set A) : 1,969 synthetic sketches generated through the XDoG algorithm from Danbooru2020 [41]. (Test Set B) : 2,778 real sketches collected by [4]. (Test Set C) : 1,194 real sketches collected by ourselves from the Internet.

B. Experimental Settings

For both training and testing, all images are resized to 256×256 pixels using mirror padding. To mitigate overfitting concerns arising from synthetic sketch extraction, we applied sketch simplification [43] to all sketches. An illustration of this image pre-processing is presented in Figure 4. Our model

Fig. 5. Qualitative examples of our method and the state-of-the-art non-reference-based sketch colorization methods. From left to right: (a) Line Art, (b) Tanpopo [15], (c) Satsuki [15], (d) Canna [15], (e) CCGAN [9], (f) AlacGAN [4], (g) DPGAN [5], (h) Style2Paints, (i) ControlNet [18], (j) Ours, and (k) Ground Truth.

is trained using a batch size of 8, and the default learning rate is set to 0.002, with no weight decay. In terms of optimization, we employ the Adam optimizer [44] with $\beta_1 = 0$ and $\beta_2 = 0.99$. All experiments are conducted on four Nvidia GTX 1080Ti GPUs.

C. Overall Performance

We conduct a comparative analysis of our method against several state-of-the-art sketch colorization techniques, which include Tanpopo [15], Satsuki [15], Canna [15], CCGAN [9], AlacGAN [4], DPGAN [5], Style2paints,[2] CWR [14], and ControlNet [18]. Furthermore, we compare our results with several recently proposed photorealistic image colorization methods to highlight the superiority of our method in terms of sketch colorization, including GCPGAN [35] and Deoldify.[3] For the purpose of comparison, all existing methods are trained and evaluated using our datasets. To ensure fairness in the comparison, we utilized the default descriptor settings for the textual input of ControlNet [18]. We included positive

[2]https://github.com/lllyasviel/style2paints
[3]https://github.com/jantic/DeOldify

TABLE I
PERFORMANCE COMPARISONS WITH STATE-OF-THE-ART METHODS. THE BEST RESULTS ARE IN **BOLD**

Test Set	Metrics	Gray-Scale Methods		Sketch Methods									
		Deoldify	GCPGAN	Canna	Tanpopo	Satsuki	CCGAN	AlacGAN	DPGAN	Style2Paints	CWR	ControlNet	Ours
A	FID ↓	64.7	40.6	108.7	81.3	125.8	45.3	53.4	52.3	83.7	44.6	69.2	**29.3**
	MOS ↑	2.9	2.3	3.2	2.5	2.4	2.9	2.7	2.4	3.0	3.1	3.3	**3.4**
B	FID ↓	76.5	43.0	103.2	85.4	81.9	45.0	52.3	50.4	70.1	39.2	83.5	**25.5**
	MOS ↑	3.1	3.0	3.1	2.5	2.6	3.0	2.5	2.4	3.1	3.2	3.5	**3.5**
C	FID ↓	104.1	67.8	139.0	106.3	113.1	73.7	84.4	80.3	102.3	59.4	69.0	**47.1**
	MOS ↑	2.5	2.7	2.5	2.7	2.4	2.9	2.8	2.1	3.0	3.1	3.1	**3.2**

descriptors such as "best quality" and "extremely detailed" were used, and negative descriptors such as "longbody," "lowres," "bad anatomy," "bad hands," "missing fingers," "extra digit," "fewer digits," "cropped," "worst quality," and "low quality".

1) Quantitative Evaluation: Unlike photo-realistic image colorization, assessing sketch colorization quality is subjective. Aligning with prior studies [4], [5], [9], [14], we utilize Fréchet Inception Distance (FID) [45] for quantitative analysis, where a lower FID indicates a distribution closer to real color illustrations, suggesting enhanced diversity and quality. However, since FID only evaluates certain aspects of image quality, we also conduct a Mean Opinion Score (MOS) user study to reflect human perceptual judgments of visual quality and to compare our method with others. Sixty sketches were chosen and colorized by each method. Twenty volunteers rated the colorization results from 1 (poor) to 5 (excellent), focusing on, diversity, and visible artifacts like color bleeding and semantic inconsistencies. The average of these scores is calculated as the MOS for each method.

The quantitative results are provided in Table I. Our method is observed to yield the lowest FID score on both synthetic sketches (Test Set A) and real sketches (Test Sets B & C), indicating that the images generated by our method have a distribution closer to real color illustrations compared to those produced by other methods. It is noteworthy that ControlNet's performance relies on the generative capability of large models and, as such, is also impacted by the data distribution inherent to these large models. Therefore, even after being trained on our dataset, the results produced by ControlNet may exhibit a higher FID score and an MOS comparable to our method. Additionally, our method attains the highest MOS, signifying that, compared to state-of-the-art methods, our model generates the most perceptually pleasing colorization results with higher diversity and fewer visible artifacts.

2) Qualitative Evaluation: Figure 5 displays qualitative examples to visually compare our method with other state-of-the-art, non-reference-based sketch colorization methods. It is evident that the colorization results of existing methods exhibit similar and restricted color patterns in each image (best viewed in each row), whereas our method introduces vivid and diverse color styles across different examples. Furthermore, our method effectively mitigates the artifacts observed in colorization results. For instance, Tanpopo, Satsuki, and Canna generate images exhibiting limited color patterns and consequently receive higher FIDs. DPGAN tends to produce noticeable color bleeding effects, particularly in background or hair regions. CCGAN struggles with semantic color inconsistency across various regions, such as clothes or hair, and tends to colorize all sketches in similar styles. ControlNet, leveraging a pre-trained large-scale model for text-to-image generation, is inherently dependent on text references. When utilizing default text prompts, it is observed that the images generated often exhibit analogous color styles. Additionally, the technique used by ControlNet may produce results that are incongruent with the original line art; for example, in the sixth and seventh column of the $i-th$ rows, the generated backgrounds, the open eyes in the last two columns, and the extra fingers are inconsistent with the information depicted in the original line art. In contrast, our proposed method yields more perceptually satisfying and cohesive results.

Figure 6 presents a visual comparison with photo-realistic image colorization methods. It is important to note that the characteristics inherent to natural image colorization and anime sketch colorization can significantly differ. Photo-realistic grayscale images encompass rich structural details, and there are no significant, explicit region divisions. Therefore, conventional convolutional structures are able to exploit these intricate patterns and successfully execute colorization in a pixel-wise manner. However, the line strokes delineate the entire sketch into several explicit regions, each region typically maintaining its own pattern. Hence, it is observable that these photo-realistic image colorization methods struggle to learn a promising color style for each region, resulting in numerous artifacts. In contrast, our RASC adeptly captures structural features in a region-wise manner, which incorporates the local style and the global dependency of each region, yielding fewer artifacts.

D. Ablation Study

1) Ablation Study for Region Guidance: Initially, we contrast our comprehensive model with the baseline architecture, which does not incorporate region maps during the colorization process (refer to the architecture in Figure 2 excluding Region Map and Region Formulation). As observed in Table II, the absence of the region map in our method results in more conspicuous artifacts, thereby accruing a higher average FID score of 36.59 across all test sets. This score signifies inferior overall quality compared to our full model's score of 33.95. Additionally, we constructed a simplified architecture where only the region map is utilized, without implementing the region formulation. In this architecture, the input is received as the concatenation of line art and the region map in the encoder, and the encoder output, $\mathbf{X}_m^e$, is subsequently concatenated in the decoder. As indicated in the penultimate row of Table II, this simplified architecture achieves an FID score of 36.32, surpassing our baseline but not

Fig. 6. Qualitative examples of our method and the state-of-the-art methods used for grayscale image colorization. From top to bottom: (a) input sketch, (b) Deoldify, (c) GCPGAN [35], (d) ours, and (e) Ground Truth.

Fig. 7. Visual comparisons for CWR [14] and ablation study of our method. From left to right: (a) Sketch, (b) CWR [14] (c) Ours w/o region maps (d) Ours w/o region formulation (e) Ours w/ region guidance. For each colorization result, we have incorporated zoomed-in versions of the marked regions.

reaching the performance of our complete architecture. This emphasizes the effectiveness of integrating region maps. All implementations incorporating region guidance (region map and formulation), referenced in the bottom row of Table II, register the lowest FID score, attesting to the potency of our Region Formulation Module. Considering that FID only evaluates the overall quality and cannot accurately represent

TABLE II

AVERAGE EVALUATION RESULTS FOR CWR [14] AND DIFFERENT VARIANTS OF OUR APPROACH ON THREE TEST SETS

Method	FID ↓	diversity ↑	fewer artifacts ↑
CWR [14]	47.69	2.86	2.56
Ours w/o region maps	36.59	3.34	2.02
Ours w/o region formulation	36.32	3.38	2.72
Ours w/ all region guidance	**33.95**	**3.43**	**3.08**

human perceptual judgments, we conduct an additional user study to assess the following two aspects of the colorization results: (1) *fewer artifacts* such as color bleeding and semantic color inconsistency, and (2) the *diversity* of generated color styles and patterns. The average MOS score of all volunteers is reported in Table II. It is evident that, in the absence of region guidance, our model maintains substantial diversity but scored significantly lower on the criterion of *fewer artifacts*. This provides strong evidence that the inclusion of region maps and region modulation, as proposed in our RASC method, does indeed enhance both diversity and the reduction of artifacts in the colorization results.

Figure 7 presents the visual comparison results. In the absence of region maps' guidance, our method often exhibits noticeable artifacts in regions within the same semantic context. For example, in the first, third, and fourth rows, it produces obvious color gradients and bleeding colors in the region of the arms, legs, and shoulders. Also, it colorizes different colors for the two eyes of the character in the second row, leading to semantic inconsistency. Without the region formulation (solely using the region map), it prevents artifacts to some extent since it contains the region maps of the corresponding sketch. However, semantic inconsistency still exists in the region of arms, and eyes in the first, and second rows. In contrast, our full method manages to reduce color bleeding effects (*e.g.*, shoulder in the first row, rhombus patches in the third row, and sleeve in the fourth row) by learning the consistent style for each region under the guidance of region maps. Also, our full model leverages consistent color for regions that have the same semantic meanings (*e.g.*, eyes in the second row) since it fully exploits the structural

Fig. 8. Upper Segment: Illustration of multi-colorization results for one sketch. Lower Segment: Illustration of the interpolation results between two different noise codes $\mathbf{w}_1, \mathbf{w}_2 \sim \mathcal{N}(0, 1)$. We can see that during the interpolation, the color styles of the images are: brown $\rightarrow$ green (brown + blue) $\rightarrow$ blue.

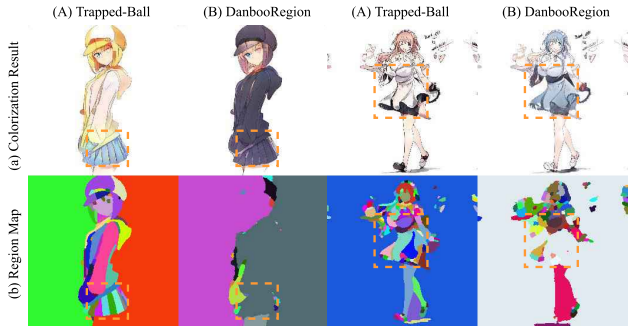


Fig. 9. Visual comparisons between trapped-ball region map formulation and danbooregion map formulation.

information by using region-wise formulation to learn the correlations among different regions.

2) Comparison With Region-Aware Architecture CWR: We further validate the superiority of our region map and region-aware architecture by contrasting it with CWR [14], which integrates skeleton region maps from DanbooRegion [23] for line art colorization. Figure 7 illustrates the respective visual comparative results. It is observable that CWR continues to manifest relatively conspicuous artifacts in regions with analogous semantic meanings, whereas our method produces visually pleasing results with fewer visual artifacts. It should be noted that CWR employs a skeleton map to furnish region guidance for colorization, which can indeed mitigate artifacts to a degree as it implicitly encompasses the region structural information of the corresponding sketch. Nonetheless, the absence of explicit constraints prevents the network from learning a consistent style for each region, as both skeleton maps and sketches are processed purely in a pixel-wise manner. Consequently, the network remains unable to fully leverage region-wise features derived from region information, resulting in visual artifacts, including color bleeding effects.

In contrast, our approach employs explicit region partitions via the region map, subsequently aggregates region-wise information through the feature aggregation process (Section III-B.1), and discerns local styles and global correlations at the regional level rather than the pixel level through the region correlation process (Section III-B.2). Consequently, our model is proficient in maximizing the utility of region information, yielding visually appealing colorization results. As demonstrated in Table II, it is evident that our method attains markedly lower FID and superior user study scores in comparison to CWR, thereby confirming the superiority of our region map and region-aware architecture.

E. Controllable Multi-Colorization for One Sketch

Our RASC accomplishes style-content disentanglement. In RASC, the latent code $\mathbf{w}$ predominantly governs the global style of colorization, while the multi-level feature maps, integrated with region-based modulation, essentially influence the local variations and structure-correlated contents. Therefore, our RASC is endowed with the capability to yield multiple colorization results, each reflecting varied color styles for a singular sketch, all while preserving the structural information inherent in the sketch. To elaborate further, during the inference stage, rather than employing a latent code as predicted by the MLP layer, we can initialize noise $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ randomly, serving as a global style for a given sketch, and continue to utilize the encoder Enc to extract content information from the sketch. Consequently, numerous colorization results, each showcasing varied color styles, can be achieved. As depicted in the upper segment of Figure 8, a range of images, each exhibiting distinct color styles, is generated for a single provided sketch. Moreover, our method also leverages the advantages of the interpretable controls inherent in GANs, enabling the achievement of controllable and smooth transitions through traversing the latent space of $\mathbf{w}$. More precisely, by performing linear interpolation between two distinct noise codes $\mathbf{w}_1, \mathbf{w}_2 \sim \mathcal{N}(0, 1)$, we can achieve a smooth transition between their respective colorization results. As illustrated in the lower segment of Figure 8, the color style of the images transitions seamlessly from brown to a blend of brown and blue, and ultimately to blue.

F. Evaluations on Different Types of Region Maps

In this section, we analyze the impact of employing different region maps from the Trapped-Ball algorithm [24] and DanbooRegion [23] on our RASC architecture. The quantitative evaluations, represented in Table III, reveal that the Trapped-Ball algorithm, although more time-consuming, generates marginally better results compared to DanbooRegion, owing to its traditional algorithms which perform fine-grained region divisions based on the structural sketch. In contrast, DanbooRegion, utilizing a network to learn the transformation from sketches to regions directly, often compromises accuracy. This discrepancy is evident in Figure 9, where the fine-grain and comprehensiveness of region maps created by the Trapped-Ball algorithm enhance the performance of our model compared to those generated by DanbooRegion.

Experiments were performed to assess the impact of employing fixed-size patches as substitutes for region maps

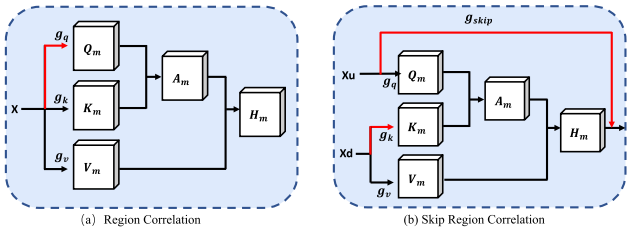
Fig. 10. Visual comparisons between fixed-sized region map formulation and our trapped-ball region map formulation.

TABLE III

AVERAGE EVALUATIONS OF REGION MAP FORMULATIONS OBTAINED BY DIFFERENT METHODS ON THREE TEST SETS

Method	FID ↓	MOS ↑	sec./img ↓
Trapped-Ball [24]	**33.95**	**3.34**	2.53
DanbooRegion [23]	35.43	3.24	**0.05**

and to investigate potential fluctuations in model efficacy. The average number of patches in the region maps, based on training dataset statistics, was 166, with a maximum of 729. We chose five patch numbers: 1, 64, 169, 256, and 1024, each corresponding to a unique patch size (Figure 10). When the patch size is equal to the input image size of $256 * 256$, the produced image displays uniform style and pattern due to our region formulation mechanism's feature aggregation (Figure 10, first column). An increase in patch numbers is correlated with enhanced diversity in the generated images, as illustrated in Table IV, due to the varied local region information captured by the patches. Additionally, an increase in patch numbers amplifies correlation learning between patches, thus improving global dependency and mitigating colorization artifacts. Notably, when the patch size is set to $8 * 8$, following three downsampling instances, the pixel-level self-attention operation is indeed performed on the image. Although improvements are evident across all metrics, the network is yet to fully leverage the region-wise feature



(a) Region Correlation (b) Skip Region Correlation

Fig. 11. Two region correlation structures for investigating gradient conflict. The red line represents that the gradient is positively correlated with the total gradient of the branch.

TABLE IV

AVERAGE EVALUATIONS ON DIFFERENT FIXED-SIZE PATCHES OF REGION MAP FORMULATIONS ON BOTH THREE TEST SETS

Patch type	FID ↓	diversity ↑	fewer artifacts ↑
$256 * 256$	37.60	2.40	2.43
$32 * 32$	35.12	2.73	2.53
$20 * 20$	35.61	2.87	2.64
$16 * 16$	35.02	2.84	2.76
$8 * 8$	35.00	3.03	2.88
Trapped-Ball Region [24]	**33.95**	**3.43**	**3.08**

based on region map, resulting in color bleeding as depicted in the fifth column of Figure 10. Our method fully considers the individual region feature and the correlation between regions, achieving the improvement of diversity and fewer artifacts as shown in Table IV

G. Exploring Gradient Conflicts in Our Region Correlation

In our study, we conducted an investigation into the gradient conflict problem [25] associated with our Region Correlation. To achieve this, we implemented a visualization of the gradient flow throughout the attention module. This approach facilitated a thorough examination of the contribution made by each individual gradient branch and the total gradient, resulting in a comprehensive understanding of their respective influences on the model's learning process. Figure 11(a) illustrates the region correlation structure within our RASC, specifically depicting the component of the Region Correlation. Herein, the three branches are denoted as g_q, g_k, and g_v, with the total gradient represented as $g_q + g_k + g_v$. We separately calculate the cosine values for each branch: $cos(g_q, g_q + g_k + g_v)$, $cos(g_k, g_q + g_k + g_v)$, and $cos(g_v, g_q + g_k + g_v)$, symbolizing the cosine similarity values between the individual gradient branches and the total gradient. This analysis aids in identifying the degree of correlation of each gradient with the total gradient, providing insights into the gradient flow throughout the learning process.

In Figure 12, each row is associated with a distinct random seed, and each column illustrates the similarity of gradients. The uniformity in histogram distributions across varying random seeds substantiates the robustness of our experimental outcomes. For each distribution, the x-axis delineates the similarity value, the y-axis denotes the epoch, and the z-axis quantifies the number of statistical points. Due to the complete positive correlation of g_q and partial positive correlation of g_k, we cannot eliminate the gradient of g_q and g_k in our module, unlike what is done in SGA [25]. To further investigate whether conflicts exist in g_q and g_k within the self-attention mechanism, and can we use SGA to improve the capability of our model by stopping the gradients? We attempt to add skip

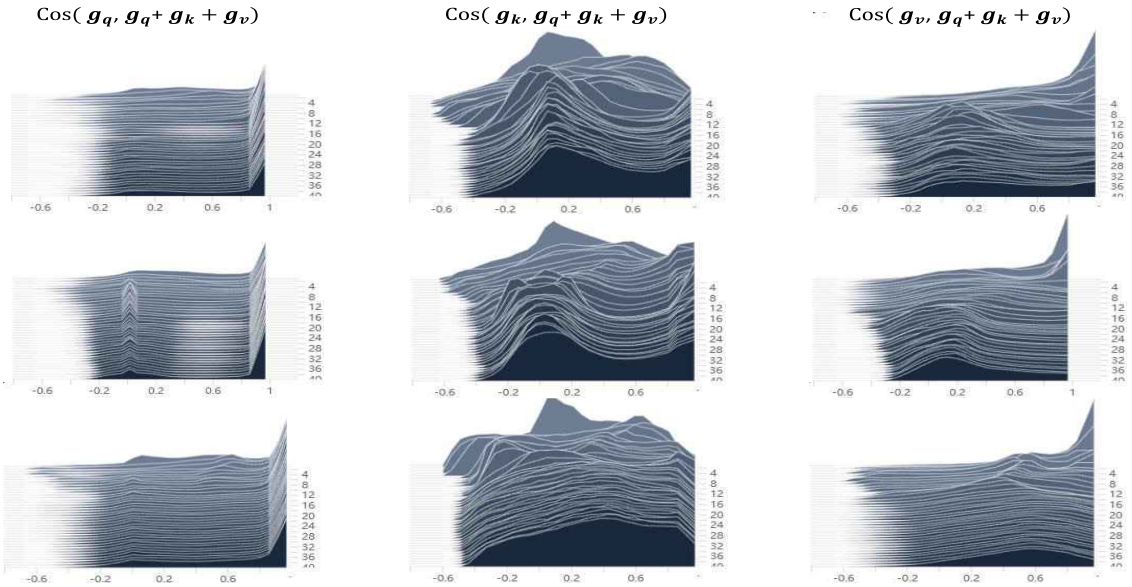


Fig. 12. The histograms of the gradient cosine value distribution in region correlation.

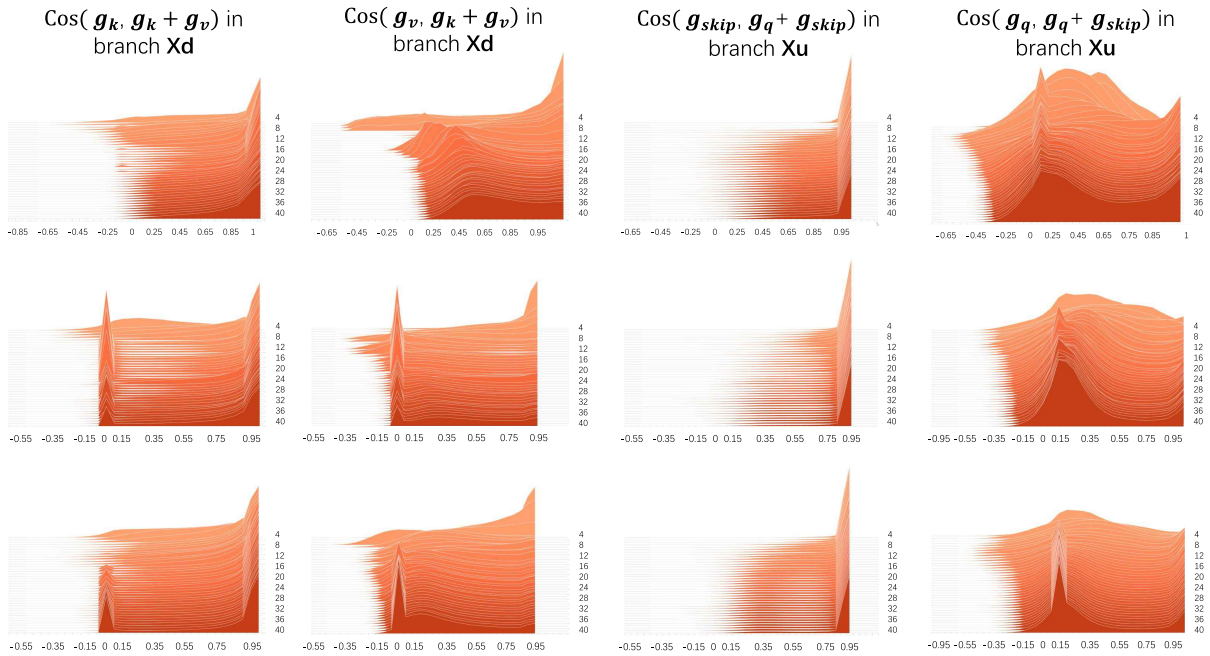


Fig. 13. The histograms of the gradient cosine value distribution in skip region correlation.

connections in our region correlation, obtained Skip Region Correlation, which is followed by the operation of SGA [25] as shown in Fig.11(b). The results of visualizing the gradient flow as shown in Figure 13, during the training process, g_q and g_v represent the problem of gradient conflict. Specifically, the g_k and g_v branches compete with each other to obtain the dominant position, and as the number of training increases, g_k occupies the dominant direction, while in g_{skip} and g_q, g_{skip} always occupies the dominant direction. In both branches, g_q and g_v exhibit a significant number of statistical points with $cos(,) \leq 0$, indicating the presence of conflicts. Hence, the SGA cannot be used in our network because there is no gradient update multiplication between g_q and g_v. Besides, we observe that in all histograms, there is always a half of gradient points that are positively correlated with the total gradient of the respective branch ($\cos(g, g + \mathbf{g}_q) > 0$). Therefore, directly applying a strategy to stop gradients in g_q or g_v also requires careful consideration. We give a simple example to illustrate the problem.

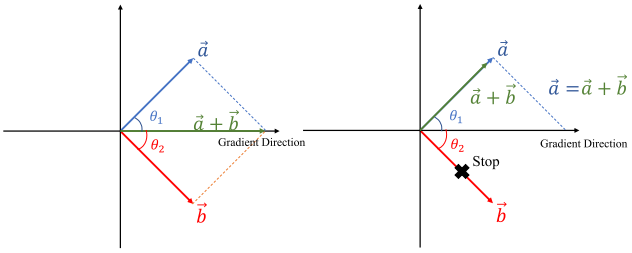


Fig. 14. Deviation in gradient direction after stopping gradient on **b**.

Fig. 15. Illustration of our proposed method's limitations. It is challenging to colorize sketches with highly complicated regions.

Let's denote two gradients, $\vec{a}$ and $\vec{b}$, both from the same branch. The total gradient is represented by a green line on the abscissa axis (as seen on the left in Fig. 14). Suppose the angle θ_1, $\theta_2 = 45$ degrees, the approximate cosine values in this scenario would be $\cos(\vec{a}, \vec{a}+\vec{b}) \approx 0.525$ and $\cos(\vec{b}, \vec{a}+\vec{b}) \approx 0.525$, indicating a positive correlation with the total gradient for both gradients. If the gradient on $\vec{b}$ is stopped, the total gradient becomes just $\vec{a}$, altering the original total gradient direction, and impacting the network's update direction. When both $\vec{a}$ and $\vec{b}$ are positive, stopping $\vec{b}$ can cause the overall gradient direction to deviate from the original direction to $\vec{a}$ (as seen on the right in Fig. 14). Addressing this deviation, especially for negatively correlated gradients, is considered important for future work.

V. Conclusion and Limitations

In this paper, we present an innovative model with Region Assist Sketch Colorization (RASC) for non-reference-based sketch colorization, aiming to produce diverse and realistic color patterns. Our model incorporates an additional region map, fully leveraged by a specially designed region-aware architecture to enhance the quality of colorization results. Specifically, we propose a unique hierarchical Region-based Modulation (RM) Block, developed to formulate regional patterns using the region feature aggregation, region correlation, and region feature broadcast module. Unlike preceding studies, our proposed model processes sketch features explicitly in a region-wise manner rather than on a pixel-wise basis, accurately depicting both the local style and global context of regional patterns, which in turn reduces synthesis artifacts in the colorization results. Comprehensive experiments conducted on both synthetic and real sketches demonstrate that our proposed method outperforms the state-of-the-art methods. Also, experiments demonstrated that our model can achieve controllable and smooth multi-colorization for one sketch by manipulating the latent code.

Limitation: Although our proposed method significantly alleviates the artifacts in colorization results, it may fail to produce reasonable colorization results when dealing with the inputs depicting highly complicated region divisions. As shown in Figure 15, the corresponding region maps contain a large number of small fragments due to the complicated and dense strokes in sketches. In this case, our method cannot adequately learn the style for each region and their implicit semantic correlations, and may predict the same color for nearby regions that actually represent different semantics, *e.g.*, sleeves, skirts, and head-wears. In future work, we will improve the architecture for scenarios containing highly complicated regions.

References

[1] Y. Qu, T.-T. Wong, and P.-A. Heng, "Manga colorization," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 1214–1220, Jul. 2006.

[2] T.-H. Sun, C.-H. Lai, S.-K. Wong, and Y.-S. Wang, "Adversarial colorization of icons based on structure and color conditions," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 683–691.

[3] L. Zhang, C. Li, T.-T. Wong, Y. Ji, and C. Liu, "Two-stage sketch colorization," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–14, Dec. 2018.

[4] Y. Ci, X. Ma, Z. Wang, H. Li, and Z. Luo, "User-guided deep anime line art colorization with conditional adversarial networks," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1536-1544.

[5] Z. Dou, N. Wang, B. Li, Z. Wang, H. Li, and B. Liu, "Dual color space guided sketch colorization," *IEEE Trans. Image Process.*, vol. 30, pp. 7292–7304, 2021.

[6] J. Lian and J. Cui, "Anime style transfer with spatially-adaptive normalization," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[7] A. Maejima, H. Kubo, T. Funatomi, T. Yotsukura, S. Nakamura, and Y. Mukaigawa, "Graph matching based anime colorization with multiple references," in *Proc. ACM SIGGRAPH Posters*, Jul. 2019, pp. 1–2.

[8] H. Kim, H. Y. Jhoo, E. Park, and S. Yoo, "Tag2Pix: Line art colorization using text tag with SECat and changing loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9055–9064.

[9] G. Zhang, M. Qu, Y. Jin, and Q. Song, "Colorization for anime sketches with cycle consistent adversarial network," *Int. J. Performability Eng.*, vol. 15, no. 3, pp. 910–918, 2019.

[10] C. W. Seo and Y. Seo, "Seg2pix: Few shot training line art colorization with segmented image data," *Appl. Sci.*, vol. 11, no. 4, p. 1464, Feb. 2021.

[11] Y. Zhao, L.-M. Po, K.-W. Cheung, W.-Y. Yu, and Y. A. U. Rehman, "SCGAN: Saliency map-guided colorization with generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3062–3077, Aug. 2021.

[12] J. Lee, E. Kim, Y. Lee, D. Kim, J. Chang, and J. Choo, "Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5800–5809.

[13] L. Zhang, C. Li, E. Simo-Serra, Y. Ji, T.-T. Wong, and C. Liu, "User-guided line art flat filling with split filling mechanism," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9884–9893.

[14] R. Cao, H. Mo, and C. Gao, "Line art colorization based on explicit region segmentation," *Comput. Graph. Forum*, vol. 40, no. 7, pp. 1–10, Oct. 2021.

[15] (2017). *Petalica Paint*. [Online]. Available: https://petalica-paint.pixiv.dev/index_en.html

[16] J. Huang, J. Liao, and S. Kwong, "Semantic example guided image-to-image translation," *IEEE Trans. Multimedia*, vol. 23, pp. 1654–1665, 2021.

[17] T.-T. Fang, D. M. Vo, A. Sugimoto, and S.-H. Lai, "Stylized-colorization for line arts," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 2033–2040.

[18] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3836–3847.

[19] C. Mou et al., "T2I-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," 2023, *arXiv:2302.08453*.

[20] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[23] L. Zhang, Y. Ji, and C. Liu, "DanbooRegion: An illustration region dataset," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 137–154.
[24] S.-H. Zhang, T. Chen, Y.-F. Zhang, S.-M. Hu, and R. R. Martin, "Vectorizing cartoon animations," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 4, pp. 618–629, Jul. 2009.
[25] Z. Li, Z. Geng, Z. Kang, W. Chen, and Y. Yang, "Eliminating gradient conflict in reference-based line-art colorization," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 579–596.
[26] X. Liu et al., "Intrinsic colorization," in *Proc. ACM SIGGRAPH Asia Papers*, Dec. 2008, pp. 1–9.
[27] B. Li, Y.-K. Lai, M. John, and P. L. Rosin, "Automatic example-based image colorization using location-aware cross-scale matching," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4606–4619, Sep. 2019.
[28] G. Charpiat, M. Hofmann, and B. Schölkopf, "Automatic image colorization via multimodal predictions," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2008, pp. 126–139.
[29] A. Y.-S. Chia et al., "Semantic colorization with internet images," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 1–8, Dec. 2011.
[30] R. K. Gupta, A. Y.-S. Chia, D. Rajan, E. S. Ng, and H. Zhiyong, "Image colorization using similar images," in *Proc. 20th ACM Int. Conf. Multimedia*, Oct. 2012, pp. 369–378.
[31] S. H. Kang and R. March, "Variational models for image colorization via chromaticity and brightness decomposition," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2251–2261, Sep. 2007.
[32] A. Bugeau, V.-T. Ta, and N. Papadakis, "Variational exemplar-based image colorization," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 298–307, Jan. 2014.
[33] Z. Cheng, Q. Yang, and B. Sheng, "Colorization using neural network ensemble," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5491–5505, Nov. 2017.
[34] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware image colorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7965–7974.
[35] Y. Wu, X. Wang, Y. Li, H. Zhang, X. Zhao, and Y. Shan, "Towards vivid and diverse image colorization with generative color prior," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14357–14366.
[36] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018, *arXiv:1809.11096*.
[37] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.
[38] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[39] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., 2016, pp. 694–711.
[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
[41] G. Branwen. (2021). *Danbooru2020: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset*. [Online]. Available: https://www.gwern.net/Danbooru2020
[42] H. Winnemöller, J. E. Kyprianidis, and S. C. Olsen, "XDoG: An eXtended difference-of-Gaussians compendium including advanced image stylization," *Comput. Graph.*, vol. 36, no. 6, pp. 740–753, Oct. 2012.
[43] E. Simo-Serra, S. Iizuka, and H. Ishikawa, "Mastering sketching: Adversarial augmentation for structured prediction," *ACM Trans. Graph.*, vol. 37, no. 1, pp. 1–13, Feb. 2018.
[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, vol. 5, no. 6, 2015.
[45] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," 2017, *arXiv:1706.08500*.

Ning Wang received the B.Sc. degree in software engineering from the Dalian University of Technology, China, where she is currently pursuing the Ph.D. degree in software engineering. Her research interests include image colorization, video colorization, image processing, and multi-modal fusion.

Muyao Niu received the B.Eng. degree from the International School of Information Science and Engineering, Dalian University of Technology. He is currently pursuing the M.Sc. degree with the Graduate School of Information Science and Technology, The University of Tokyo. His research interests include computational photography, 3D vision, and computer vision.

Zhihui Wang (Member, IEEE) received the Ph.D. degree from the Dalian University of Technology, China. She is currently a Professor with the International School of Information Science and Engineering, Dalian University of Technology. She has published papers in international and national journals, and international conferences. Her research interests include pattern recognition, computer vision, and machine learning.

Kun Hu received the B.Sc. and B.Eng. degrees from Shandong University, Jinan, China, in 2013, the M.Sc. degree from Shanghai Jiao Tong University, Shanghai, China, in 2016, and the Ph.D. degree from The University of Sydney, Sydney, NSW, Australia, in 2022. He is currently a Postdoctoral Research Associate with the School of Computer Science, The University of Sydney. His research interests include pattern recognition, computer vision, and optimization.

Bin Liu (Member, IEEE) received the Ph.D. degree from the Dalian University of Technology, China. He is currently a Professor with the International School of Information Science and Engineering, Dalian University of Technology. His research interests include computer vision, medical imaging processing, three-dimensional reconstruction, and computer graphics.

Zhiyong Wang (Member, IEEE) received the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong. He is currently an Associate Professor and an Associate Director of the Multimedia Laboratory, School of Computer Science, The University of Sydney, Australia. His research interests include multimedia computing, including multimedia information processing, retrieval and management, internet-based multimedia data mining, human-centered multimedia computing, and pattern recognition.

Haojie Li (Member, IEEE) received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, China. He is currently a Professor with the International School of Information Science and Engineering, Dalian University of Technology, China. He has published papers in international and national journals, and international conferences. His research interests include pattern recognition, computer vision, and machine learning.