electronics

MDPI

Article

Thangka Sketch Colorization Based on Multi-Level Adaptive-Instance-Normalized Color Fusion and Skip Connection Attention

Hang Li [1], Jie Fang [1], Ying Jia [1], Liqi Ji [1], Xin Chen [1] and Nianyi Wang [2,*]

[1] Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou 730000, China
[2] School of Mathematics and Computer Science, Northwest Minzu Univsersity, Lanzhou 730000, China
* Correspondence: livingsailor@gmail.com

Abstract: Thangka is an important intangible cultural heritage of Tibet. Due to the complexity, and time-consuming nature of the Thangka painting technique, this technique is currently facing the risk of being lost. It is important to preserve the art of Thangka through digital painting methods. Machine learning-based auto-sketch colorization is one of the vital steps for digital Thangka painting. However, existing learning-based sketch colorization methods face two challenges in solving the problem of colorizing Thangka: (1) the extremely rich colors of the Thangka make it difficult to color accurately with existing algorithms, and (2) the line density of the Thangka brings extreme challenges for algorithms to define what semantic information the lines imply. To resolve these problems, we propose a Thangka sketch colorization method based on multi-level adaptive-instance-normalized color fusion (MACF) and skip connection attention (SCA). The proposed method consists of two parts: (1) a multi-level adaptive-instance-normalized color fusion (MACF) to fuse sketch feature and color feature; and (2) a skip connection attention (SCA) mechanism to distinguish the semantic information implied by the sketch lines. Experiments on colorizing Thangka sketches show that our method works well on two small datasets—the Danbooru 2019 dataset and the Thangka dataset. Our approach can generate exquisite Thangka.

Keywords: Thangka; machine learning; attention

Citation: Li, H.; Fang, J.; Jia, Y.; Ji, L.; Chen, X.; Wang, N. Thangka Sketch Colorization Based on Multi-Level Adaptive-Instance-Normalized Color Fusion and Skip Connection Attention. *Electronics* **2023**, *12*, 1745. https://doi.org/10.3390/electronics12071745

Academic Editor: Silvia Liberata Ullo

Received: 5 March 2023
Revised: 1 April 2023
Accepted: 4 April 2023
Published: 6 April 2023

1. Introduction

As a kind of Tibetan encyclopedia [1], Thangka art is one of Tibet's most valuable cultural heritages and one of the most precious materials for studying Tibetan history. It takes a professional Thangka painter dozens of days or even years to paint a beautiful Thangka. Thangka colorization is one of the essential parts of the Thangka painting process, which requires a lot of time and effort for professional Thangka painters. Recently, the colorization of a given image has attracted much attention in computer vision. The machine learning-based sketch colorization method enables us to use digital means to create and preserve Thangka better.

Although great progress [2,3] has been made in sketch colorization methods, there is no available solution for the task of Thangka sketch colorization due to three main reasons: (1) Thangka artworks are extremely colorful, which makes both traditional non-learning methods and standard convolution-based learning methods hard to color correctly; (2) the line density of the Thangka makes it difficult for the existing methods to correctly define what semantic information the Thangka lines imply; and (3) the existing Thangka dataset is too small to be trained well with the existing colorization methods.

Reference-based sketch image colorization [2–4] has excellent potential for sketch colorization of Thangka, and these related works have achieved remarkable results. However,

these existing reference-based sketch image colorization methods are still unable to produce satisfactory colorization results because of two challenges:

Challenge 1: Existing reference-based sketch image colorization algorithms cannot correctly extract the color features of Thangka. Compared with the existing datasets, such as face images, landscapes, indoor scenes, flowers, animals, animation and cars, etc., the colors in the Thangka dataset are extremely rich. The performance of Thangka painting depends on the color feature extraction. However, the existing colorization methods are mainly applied to color animation works which only contain simple colors and single structures.

Challenge 2: Compared with other sketch images, the lines of the Thangka sketches are too dense for algorithms to define what semantic information the lines imply, which leads to problems in the colorization process (for example, in Figure 1, the results of existing methods show obvious artifacts, wrong colors, and color confusion).

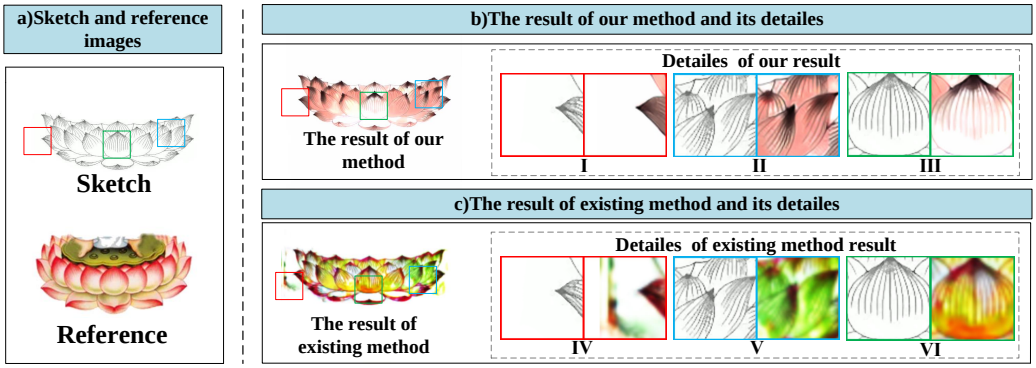

Figure 1. The proposed MACF-SCA obtains the best colorization map with more accurate colors and better visual effects (**b**). Existing reference-based methods cannot accurately migrate the color semantic information (**c**). The detailed parts show that our model can accurately distinguish the semantic information of dense lines (**I**,**II**,**III**), and the existing model shows obvious visual artifacts, color confusion and color errors (**IV**,**V**,**VI**).

In this paper, a multi-level adaptive-instance-normalized color fusion (MACF) and skip connection attention mechanism (SCA) is proposed (Figure 2) to solve the Thangka sketch colorization. The proposed method consists of two parts:

First (solution for Challenge 1), a new multi-level adaptive-instance-normalized color fusion (MACF) is proposed to fuse rich color features with sketch features efficiently. MACF consists of a combination of four identical Convolutional AdaIN ReLU (CAR) modules (Section 3.3). Firstly, color features and sketch features are extracted using a color feature encoder and sketch feature encoder, respectively. Then they are fed together into a multi-stage MACF for fusion. The multi-level MACF ensures the accurate fusion of color features and sketch features.

Second (solution for Challenge 2), a new semantic information distinction based on skip connection attention (SCA) is proposed to focus on sketch lines. SCA is extremely sensitive to subtle objects, allowing our model to accurately distinguish what semantic information the subtle lines imply. Our skip connection attention-trained Thangka colorization model not only provides accurate recognition of semantic information but also avoids overfitting.

Experiments on the colorization of Thangka sketches show that our method can generate high-quality Thangka colorization results in one step without post-processing. In summary, the main contributions of this work are:

1. We propose a multi-level color fusion module multi-level adaptive-instance-normalized color fusion (MACF), which can accurately fuse color features with sketch features and generate high-quality colorization works.

2. We propose a skip connection attention (SCA) module for accurately distinguishing semantic information consisting of dense lines.
3. For the first time, we present a framework applicable to the Thangka sketches colorization, and we also constructed a new Thangka dataset (5081 images).

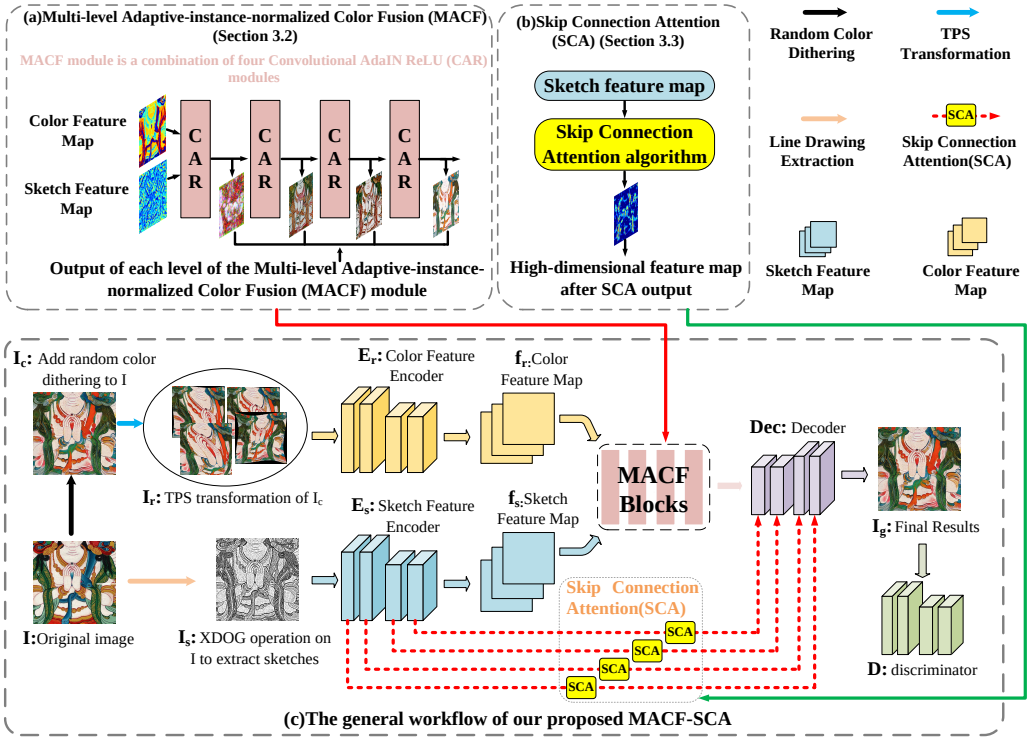

Figure 2. The proposed Thangka sketch colorization method consists of two parts: (1) a multi-level adaptive-instance-normalized color fusion (MACF) module for the fusion of color features and sketch features, and (2) the integration of a skip connection and attention (SCA) mechanism, which scales the extracted sketch features to capture a large enough receptive domain to distinguish the semantic information of the lines accurately.

2. Related Work

2.1. Automatic Sketch Colorization

Automatic sketch colorization methods based on deep learning [5–11] have received increasing attention in recent years. Relying on the powerful representation capability of deep neural networks, automatic sketch colorization methods can be implemented by designing various network structures and using large-scale image datasets. Liu et al. [5] used a feed-forward deep neural network as a generator to output color images with pixel-level resolution using sketches as the input. Frans et al. [6] proposed two tandem adversarial networks for the automatic colorization of sketch images. Recent studies [7,11] have improved using the U-net network and proposed a U-net-based architecture for automatic sketch colorization.

For all automatic sketch colorization methods, there are two main problems: (1) these methods are sensitive to visual artifacts when the sketch has complex content with multiple objects, and (2) the existing methods tend to output single-color results and have no multimodality since the network parameters are fixed.

2.2. User Prompt Based Colorization

Early interactive colorization methods [12,13] used low-similarity metrics to propagate stroke colors. Recently, some algorithms [5,14–17] introduced manual guidance to apply initial color points or strokes to the entire sketch image. Ci et al. [14] proposed a deep conditional adversarial architecture to robustly train the network to make synthetic images

more natural and realistic. Zhang et al. [15] proposed a two-stage colorization framework based on semi-automatic learning to color sketches with appropriate colors, textures, and gradients. Yuan et al. [16] proposed a tandem and U-net-based framework on a spatial attention module that can generate more consistent and higher-quality sketch colorization from the cues given by the user.

However, all these methods have limitations: (1) these palette-based colorization methods are susceptible to user aesthetic limitations, and (2) it is difficult for untrained users to select the appropriate points and associated colors from the palette.

2.3. Reference-Based Sketch Image Colorization

In contrast to the user prompt-based colorization, reference-based sketch image colorization only requires a user to select a suitable reference image based on a target sketch image. Colorization of the sketch image according to the reference style is a user-friendly method that helps a designer choose the right color image for the sketch [2–4,18–20]. With the recent rise of deep neural networks, Zhang et al. [4] integrated the residual U-net into a generative adversarial network (AC-GAN) with an auxiliary classifier for the anime sketch colorization task. Due to the limitations of the sketch-reference image pair dataset, Lee et al. [2] proposed an enhanced self-reference generation method, where the reference image is generated from the original image by color perturbation and geometric distortion, followed by an attention-based pixel feature transfer module to colorize the sketch image. Li et al. [3] proposed a stop-gradient-attention (SGA) training strategy based on [2] to eliminate gradient conflicts and help models learn better colorization correspondences.

Although these models achieved good results, the results of the Thangka sketch drawings are not satisfactory. As shown in Figure 1, a comparison between the existing method (Figure 1c) and our method (Figure 1b) shows obvious artifacts, color errors, and semantic mismatches of the existing method.

3. Methodology

The proposed sketch colorization method for Thangka consists of two parts: (1) a multi-level adaptive-instance-normalized color fusion (MACF) for fusing color features and sketch features, and (2) a skip connection attention (SCA) module that integrates skip connection and attention mechanism, which can accurately discern the semantic information of dense lines.

3.1. Overall Workflow

As shown in Figure 2, given a color image I, we first convert it to an artistic line image I_s using XDoG [21]. Then, inspired by [2], we obtain the expected colorization result I_c by adding a random color dithering on I. Next, a self-styled reference image I_r was generated by applying the thin plate spline (TPS) transformation to I_c. Finally, we use a self-supervised training process similar to [2]. In the training process, our model takes I_r and I_s as inputs and uses two independent encoders E_s (I_s) and E_r (I_r) to extract sketch features $f_s \in R_{c*h*w}$ and color features $f_r \in R_{c*h*w}$.

To carry out sketch feature alignment and color feature fusion simultaneously, the extracted color features are fused into the depth representation of the sketch through our MACF block control feature map. The final color image I_g is then generated using multiple residual blocks and a decoder with a skip connection to the sketch encoder E_s. In the end, we add an adversarial loss [22] by using a discriminator D to distinguish the output I_g and the ground truth I_c. The color style is similar to the reference image, and the content is consistent with the input sketch.

3.2. Self-Enhanced Self-Referential Learning

Due to the scarcity of the Thangka dataset, preparing reference images for Thangka sketch images and linking these two inputs for pixel-level pairing training is a crucial bottleneck; we adopt the self-reference generation method from [2]. To generate a random

reference image I_r for a given Thangka sketch I_s, we perform a spatial transformation of the original real color image I. Since I_r is essentially generated by I, this process guarantees enough color information to color I_s, which encourages the proposed model to reflect I_r in the colorization process.

Detailed information on how these conversions work is described below. First, the content transformation $C(\cdot)$ adds a specific random color perturbation on I. The resulting output C(I) is then used as the ground truth I_c for the colorization output of our model. The reason why we impose the color perturbation to the original I is to increase the training samples. The same original color image I can have different reference images. Afterward, we further apply the thin plate spline (TPS) [23] transform $T(\cdot)$, a non-linear spatial transformation operator to C(I) (or I_c), resulting in our final reference image I_r. This prevents our model from lazily bringing colors from I to the same pixel location while forcing our model to extract semantic color information only from the reference image, even if it has a different layout in space. For example, differences in orientation, shape, and posture. The above two transformations help our model learn to transfer the correct color information from the reference image to the target image.

3.3. Multi-Level Adaptive-Instance-Normalized Color Fusion (MACF)

Existing deep learning methods, such as SCFT [2] and SGA [3], have reached state-of-the-art image colorization. However, they still fail to correctly migrate rich colors through deep networks, which inevitably leads to inaccurate colorization of the final results. Our MACF consists of four identical Convolutional AdaIN ReLU (CAR) modules, and the composition of the CAR module is shown in Figure 3.

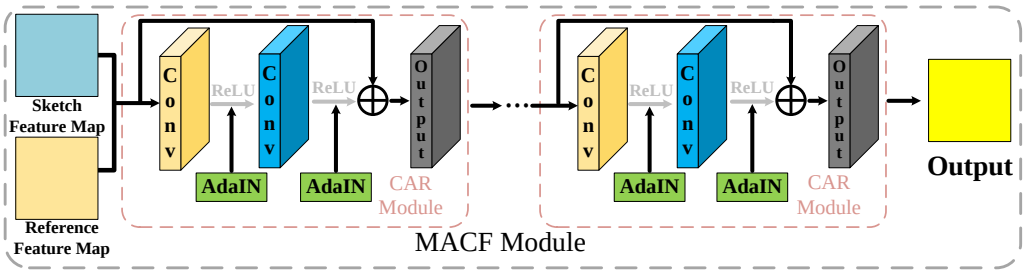

Figure 3. The proposed multi-level adaptive-instance-normalized color fusion (MACF) consists of four identical CAR modules. Given a set of sketch and reference images, different results are output by level-by-level CAR. It is a multi-level operation that enables the natural fusion of color features with sketch features.

MACF fuses the color feature map extracted by the color encoder with the sketch feature map extracted by the sketch encoder. In MACF, we use the AdaIN layer to control the input feature maps to achieve alignment of color features with sketch features. Multi-layer CAR is used to output multi-scale result maps.

3.4. Skip Connection Attention (SCA)

Inspired by [24], we note that attention gates (AG) are extremely sensitive to subtle changes, which helps us process complex textured images such as Thangka. We merged the AG into our net architecture to highlight the sketch features passing through the skip connection. For the SCA module, we provide two inputs—a complete sketch feature map and a rough feature map. As shown in Figure 2, the sketch feature information is roughly extracted for gating in the E_s encoder using the vgg19 network [25] to eliminate irrelevant noise and ambiguous responses in the skip connection. This is performed before the join operation to merge and activate only the relevant feature information in the decoder.

As shown in Figure 4, in order to accurately distinguish the semantic information of sketch lines, a sufficiently sizeable receptive field needs to be captured. When the decoder is connected to the skip connection, the attention gate calculates the activation weights in

the skip connection, locks the spatial region and scales the sketch features delivered by the skip connection. The sketch feature map is scaled using the attention factor (α) computed by AG. The spatial regions are selected by analyzing the activation and line information provided by the gating signal (G), which is collected from a coarser scale. Finally, sketch feature mapping and color feature fusion are performed in the decoder to obtain the final colorized Thangka image.

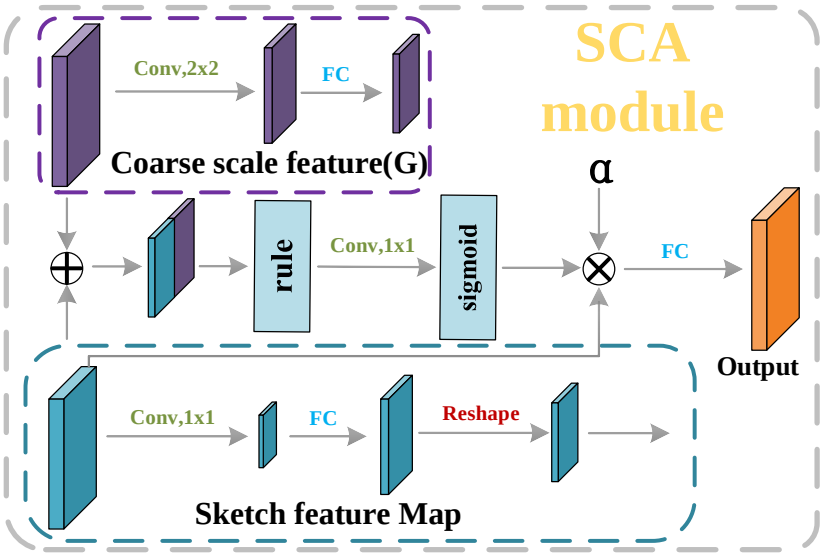

Figure 4. The proposed skip connection attention (SCA) integrates skip connection and attention mechanisms to discriminate semantic information consisting of dense lines accurately.

3.5. Loss Function

For machine learning, the design of the loss function depends on the goal of the training. In this work, the goal of Thangka sketch colorization is to give the Thangka sketch appropriate colors to show the beauty of the Thangka artwork. To achieve this goal, the Reconstruction Loss ($\mathcal{L}_{rec}$), Adversarial Gen Loss ($\mathcal{L}_{adv}$), Perceptual Loss ($\mathcal{L}_{perc}$) and Style Loss ($\mathcal{L}_{style}$) functions are used in our method.

Reconstruction Loss. According to Section 3.1, the generated image I_g and the ground truth image I_c should be stylistically consistent with the reference image I_r and retain consistent contours with the sketch image I_s, respectively. Therefore, we use the L1 [26] criterion to measure the difference between I_g and I_c, which ensures that the model adds color correctly and distinctly. The reconstruction loss can be expressed as:

$$\mathcal{L}_{rec}(G) = E_{(I_s,I_r,I_c)}[\|G(I_s, I_r) - I_c\|_1] \tag{1}$$

where G (I_s, I_r) means coloring the sketch I_s with the reference I_s and I_c is the color image.

Adversarial Gen Loss. As an adversary of the generator, the discriminator D aims to distinguish the images generated by the generator from the real ones. The output of the real/fake classifier (X) represents the probability that any image X is a real image. We chose conditional $GANs$, which uses generated samples and additional conditions [27] simultaneously. In this work, we use the input image I_s as the condition for adversarial loss because preserving the content of I_s and generating plausible fake images is important. The optimization of $D's$ loss is expressed as a standard cross entropy loss as:

$$\mathcal{L}_{adv}(G, D) = E_{(I_s,I_c)}[\log D(I_s, I_c)] + E_{(I_s,I_r)}[\log(1 - D(I_s, G(I_s, I_r)))] \tag{2}$$

where G represents the generator , D represents the discriminator, I_s is the sketch image, I_r is the reference image, I_c is the color image, and I_g is the generated image.

The first term $E_{(I_s,I_c)}[\log D(I_s, I_c)]$ represents the discriminator's loss for real images (I_s, I_c), so the goal of this term is to make the discriminator better at distinguishing real images from fake ones by making its output $\log D(I_s, I_c)$ closer to 1. The second term $E_{(I_s,I_r)}[\log(1 - D(I_s, G(I_s, I_r)))]$ represents the generator's loss for generated fake images $(I_s, G(I_s, I_r))$; this term aims to make the discriminator output $\log(1 - D(I_s, G(I_s, I_r)))$ closer to 0.

Perceptual Loss. As shown in previous work [28], perceptual loss [29] can drive the network to produce perceptually plausible outputs and has also been shown to facilitate the training of sketch colorization models [30,31]. We use the perceptual loss computed on the VGG19 network [25] pre-trained on ImageNet as the content loss of the generator as:

$$\mathcal{L}_{\text{perc}}(G) = \sum_{i=1}^{N} \frac{1}{T_i} \left[\left\| F^{(i)}(I_c) - F^{(i)}(G(I_s, I_r)) \right\|_1 \right] \tag{3}$$

where T_i is the number of elements in the i-th layer of VGG19 and $F^{(i)}$ is the feature mapping in the i-th layer.

Style loss. Lee et al. [2] have shown that style loss helps the network to produce reasonable outputs. The style loss is calculated as:

$$\mathcal{L}_{\text{style}} = E\left[\left\| \mathcal{G}\left(\phi_l(\hat{I})\right) - \mathcal{G}\left(\phi_l(I_{gt})\right) \right\|_1 \right] \tag{4}$$

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

In summary, the overall loss function for training is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{style}}\mathcal{L}_{\text{style}} \tag{5}$$

4. Experimental Results and Analysis

4.1. Dataset

We used the Danbooru 2019 and Thangka datasets to train and validate our model.

Danbooru 2019 dataset [32]. Danbooru 2019 is the most widely used dataset in animation sketch colorization. For the Danbooru 2019 dataset, we filtered 16,170 images from it for training and 2000 images for testing. It consists of objects with black background images. Since the black background has obvious area boundaries when extracting lines, we substitute all black backgrounds with white backgrounds to facilitate the extraction of sketches. This dataset is used to train our model in the cartoon domain so that the Thangka sketches have the stylistic characteristics of anime.

Thangka dataset. Since there is no publicly available Thangka dataset for our study, we collected 128 ultra-high-definition Thangka images (size 12,869 × 16,710) from the Internet and then manually cropped and cut out beautiful partial pictures of these Thangka murals containing portraits of Buddha statues, lotus bases, sacred animals, auspicious clouds, auspicious treasures, temples, etc. Finally, 5662 Thangka images (size 512 × 512) were obtained for the experiment. We allocate 5081 images for training and 581 images for testing.

To simulate the lines drawn by the artist for both the Danbooru 2019 dataset and the Thangka dataset, we used XDoG [21] to extract the sketch inputs and set the parameters of the XDoG algorithm to $\phi = 1 \times 10^9$ in order to maintain a step transition at the boundary of the sketch lines. For other parameters, we set $\sigma = 0.5$, $p = 19$, $k = 4.5$, and $\varepsilon = 0.01$ by default in XDoG.

4.2. Implementation Details

We trained our model on a single NVIDIA 3090 GPU and we set the coefficients of each loss term as follows: $\lambda_{rec} = 30$, $\lambda_{perc} = 0.01$, and $\lambda_{style} = 50$. We use the Adam solver [33] for optimization with momentum hyperparameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rates of the generator and discriminator are initially set to 0.0001 and 0.0002, respectively. For each dataset, the size of the input image is fixed as 512 × 512.

4.3. Qualitative Evaluation

We conducted experiments on the Danbooru 2019 and Thangka datasets and compared our approach with existing state-of-the-art methods that include not only reference-based line art colorization [2,3] but also image-to-image translation [20]. Figure 5 visually compares the overall qualitative results of our method with the state-of-the-art methods. Figure 5A,B show the results of the Danbooru 2019 dataset, and C,D compare the results of the Thangka dataset. The sketch and reference images are given in the first and second columns, respectively.

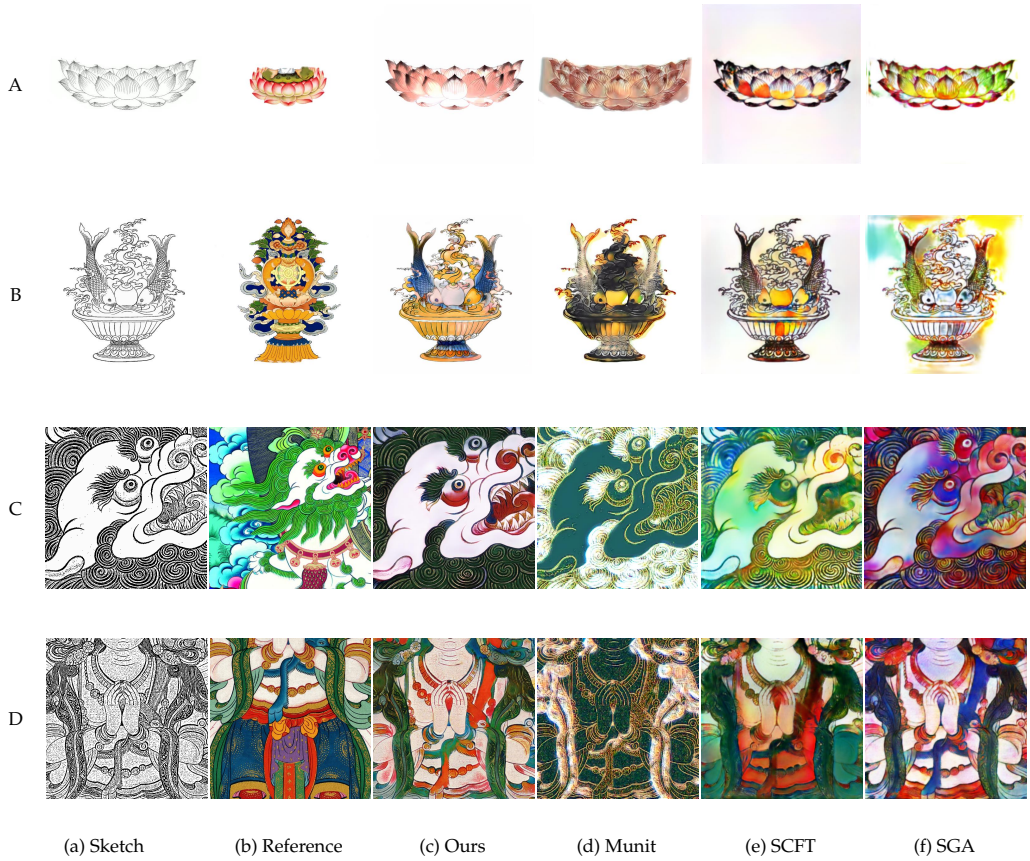

Figure 5. Colorization result of Thangka sketch. Compared with Munit [20] (**d**), SCFT [2] (**e**), and SGA [3] (**f**), our results (**c**) show correct correspondence between the sketch and the reference images. Munit [20] (**d**) does not learn semantic information correctly, SCFT [2] (**e**) shows significant artifacts, and SGA [3] (**f**) shows significant color misalignment. Column (**a**) is the sketch of the Thangka, and column (**b**) is the reference image. A,B show the results of the Danbooru 2019 dataset, and C,D compare the results of the Thangka data.

On each dataset, our model extracts the exact colors from the reference image and injects them into the corresponding positions in the sketch. For example, in the first row of Figure 5, our model colored the lotus base correctly, while SCFT [2], SGA [3] and Munit [20] all showed unsatisfactory visual effects. In contrast, our method finely fills in the same colors as the reference image. As shown in the third row of Figure 5, the results of the Thangka show that Munit [20] was unable to learn the correct semantic information of the Thangka image. SCFT [2] and SGA [3] also showed obvious color overflow, errors, and noticeable visual artifacts.

The experimental results of the Danbooru 2019 and Thangka datasets show the superiority of our method over SCFT [2], SGA [3] and Munit [20], demonstrating the advantages of our model in establishing visual correspondences and generating appropriate colors in Thangka images.

4.4. Quantitative Evaluation

In traditional sketch colorization setups, pixel-level evaluation metrics such as peak signal-to-noise ratio (PSNR) and contour retention evaluation metrics such as structural similarity index (SSIM) are widely used. The Fréchet inception distance (FID) [34] is a well-known metric used to evaluate the performance of generative models. In our study, we use the following three metrics to evaluate the results of our model quantitatively.

Fréchet inception distance (FID) [34]. FID is a well-known metric used to evaluate the performance of generative models by measuring the Wasserstein-2 distance between the feature space representation of the actual image and its generated output. A low FID score indicates that the model generates images with quality and diversity close to the real data distribution.

Peak signal-to-noise ratio (PSNR). PSNR is based on the error between corresponding pixel points and is one of the most widely used objective image evaluation metrics. A higher PSNR score indicates a better similarity between the reconstructed and ground truth color images.

Structural similarity index (SSIM). The structural similarity index, which calculates the structural similarity index (SSIM) between the reconstructed image and the original color image, measures the preservation of the contours of the drawing during the colorization process. The higher the score, the more similar the two images are; the ideal value is 1.

To evaluate the performances of different methods, we randomly selected reference and sketch images for colorization and used the above three metrics for a quantitative study. Table 1 shows the results.

Table 1. Quantitative comparisons show that our model trained for Thangka colorization outperforms the models trained by SCFT [2], SGA [3] and Munit [20] (tests are conducted on both the Danbooru 2019 dataset of 2000 images and the Thangka dataset of 581 images). FID [34] score: a lower score is better. PSNR and SSIM score: a higher score is better.

Method	Danbooru 2019			Thangka		
	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑
Munit [20]	58.38	0.68	13.83	168.89	0.19	9.09
SCFT [2]	45.60	0.80	16.23	183.65	0.48	13.26
SGA [3]	24.28	0.79	16.18	110.79	0.43	11.92
Ours	16.13	0.82	16.53	51.35	0.54	12.87

We report the FID, SSIM and PSNR scores calculated by these models on different datasets in Table 1. Our model scores show that the proposed SCA module in the model plays a valuable role in generating realistic images by establishing context-supervised semantic correspondence through skip connections. Our method produces results closest to ground truth color images, which demonstrates the realism and robustness of our method on different images. We show more examples of Thangka sketch coloring in Figure 6.

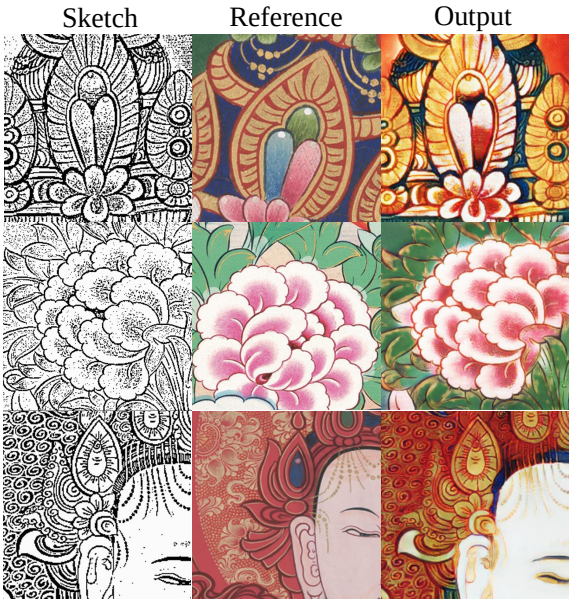

Figure 6. More colorization examples of Thangka sketches generated by our method. Our MACF-SCA generates visually excellent and semantically sound colorization result maps.

4.5. Ablation Study

We conducted several ablation experiments to validate the effectiveness of each component of our approach, namely multi-level adaptive-instance-normalized color fusion (MACF) and skip connection attention (SCA). Table 2 reports the quantitative ablation results, reflecting the validity of our model. PSNR/SSIM metrics are evaluated by paired sketch/reference inputs, and the FID is assessed by random reference.

First, we removed the MACF module to evaluate the effectiveness of multi-level adaptive-instance-normalized color fusion (MACF), which obtained poor performance in Table 2, verifying the necessity of our MACF.

Second, we conducted an ablation study on the skip connection attention (SCA) module to verify the advantages of the skip attention mechanism in our framework. Table 2 shows that the model's performance with the SCA module is significantly better than the model without the SCA module. Although a realistic image can be generated without the SCA module, it has a lower contour retention, i.e., SSIM metric.

Finally, in the third row of Table 2, we show the results of our whole model, and we can see that it performs with a significantly superior quality of image generation. The contour retention rate, i.e., the SSIM, is also the highest. Our ablation study demonstrated the effectiveness of MACF and SCA.

Table 2. The ablation study validated the effectiveness of the proposed adaptive-instance-normalized color fusion (MACF) and skip connection attention (SCA). With the presence of both MACF and SCA in the model (row 3 of Table 2), our model achieved the best results. FID [34] score: a lower score is better. PSNR and SSIM score: a higher score is better.

Method	**Danbooru 2019**			**Thangka**		
	FID↓	**SSIM↑**	**PSNR↑**	**FID↓**	**SSIM↑**	**PSNR↑**
W/O MACF	21.35	0.79	14.42	66.05	0.51	12.30
W/O SCA	18.40	0.78	15.88	60.79	0.50	12.41
FULL	16.13	0.82	16.53	51.35	0.54	12.87

We show two qualitative examples in Figure 7 that demonstrate the effectiveness of MACF and SCA.

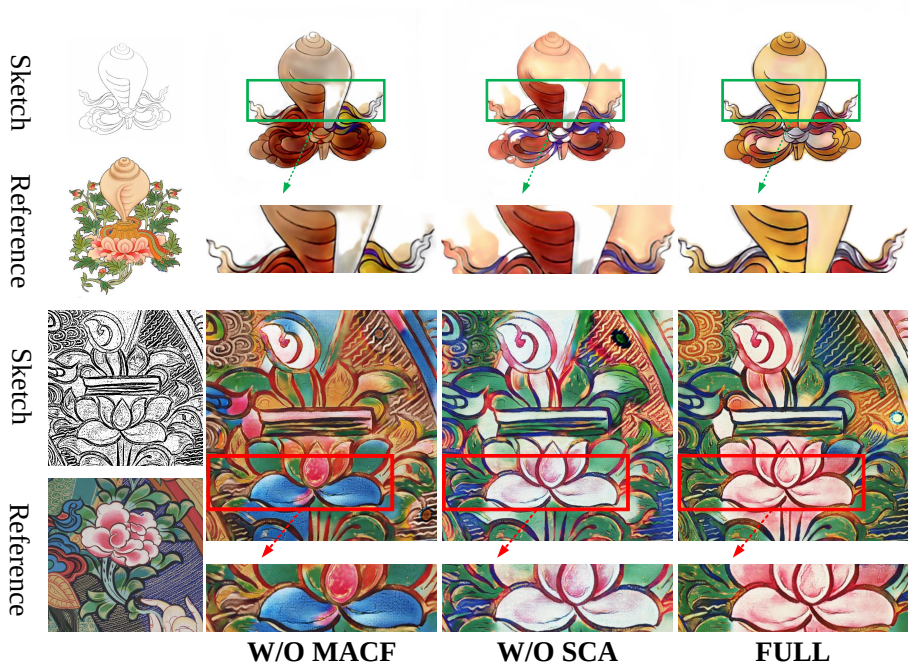

Figure 7. The proposed multi-level adaptive-instance-normalized color fusion (MACF) and skip connection attention (SCA) allow our MACF-SCA to produce visually better and more accurate colorization results by multi-level color fusion (see the fourth column). In contrast, the second column (without MACF strategy) shows obvious color confusion. In the third column (without SCA strategy), it is not difficult to find color errors due to misjudgment of line semantic information.

5. Conclusions

We propose a method of colorizing Thangka sketches based on multi-level adaptive-instance-normalized color fusion (MACF) and skip connection attention (SCA) for generating Thangka artworks. The method consists of two parts: (1) a new multi-level adaptive-instance-normalized color fusion module (MACF) for the accurate fusion of color features with sketch features, and (2) a new skip connection attention (SCA) for accurately distinguishing semantic information composed of dense lines. Experiments on two different datasets show that our method can produce more visually plausible and richer colorization maps compared to the existing methods. Both objective and subjective evaluations validated the performance of our method.

Although our method works well on anime and Thangka sketches, our outputs can still be affected by the colors and textures of style references. If the style reference contains little color information, output quality may be unsatisfactory. In addition, this study focuses more on coloring religious artworks, Thangka and anime, while other types of inputs still need further optimization and improvement. Future work includes using more related techniques such as color mapping, color gradient generation, and color blending to enhance the expressiveness and fidelity of coloring. It also includes adapting the model to accommodate more types of inputs and improving the model to handle higher resolution images.

Author Contributions: Conceptualization, H.L. and J.F.; methodology, H.L.; validation, Y.J., L.J. and X.C.; formal analysis, H.L. and N.W.; investigation, H.L. and N.W.; data curation, H.L.,Y.J., J.F., X.C. and L.J.; writing—original draft preparation, H.L.; writing—review and editing, N.W.; supervision, N.W.; project administration, N.W.; funding acquisition, N.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is jointly supported by NSFC (Grant No. 61862057, No. 62061042), and the Fundamental Research Funds for the Central Universities (Grant No. 31920210140).

Data Availability Statement: The dataset used in this study will be considered publicly available at a later stage available from livingsailor@gmail.com available upon request.

Acknowledgments: The findings and opinions expressed in this article are those of the authors only and do not necessarily reflect the views of the sponsors.

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

References

1. Wang, W.; Qian, J.; Lu, X. Research outline and progress of digital protection on thangka. In *Advanced Topics in Multimedia Research*; United Kingdom: IntechOpen 2012 ; p. 67.
2. 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, 13–19 June 2020; pp. 5801–5810.
3. Li, Z.; Geng, Z.; Kang, Z.; Chen, W.; Yang, Y. Eliminating Gradient Conflict in Reference-based Line-Art Colorization. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Part XVII, pp. 579–596.
4. 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.
5. Liu, Y.; Qin, Z.; Wan, T.; Luo, Z. Auto-painter: Cartoon image generation from sketch by using conditional Wasserstein generative adversarial networks. *Neurocomputing* **2018**, *311*, 78–87.
6. Frans, K. Outline colorization through tandem adversarial networks. *arXiv* **2017**, arXiv:1704.08834.
7. Zhang, G.; Qu, M.; Jin, Y.; Song, Q. Colorization for anime sketches with cycle-consistent adversarial network. *Int. J. Perform. Eng.* **2019**, *15*, 910.
8. Seo, C.W.; Seo, Y. Seg2pix: Few shot training line art colorization with segmented image data. *Appl. Sci.* **2021**, *11*, 1464.
9. Furusawa, C.; Kitaoka, S.; Li, M.; Odagiri, Y. Generative Probabilistic Image Colorization. *arXiv* **2021**, arXiv:2109.14518.
10. Yan, C.; Chung, J.J.Y.; Kiheon, Y.; Gingold, Y.; Adar, E.; Hong, S.R. FlatMagic: Improving Flat Colorization through AI-Driven Design for Digital Comic Professionals. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022; pp. 1–17.
11. Liu, G.; Chen, X.; Hu, Y. Anime sketch coloring with swish-gated residual U-net. In Proceedings of the Computational Intelligence and Intelligent Systems: 10th International Symposium, ISICA 2018, Jiujiang, China, 13–14 October 2018; Revised Selected Papers 10; pp. 190–204.
12. Huang, Y.C.; Tung, Y.S.; Chen, J.C.; Wang, S.W.; Wu, J.L. An adaptive edge detection based colorization algorithm and its applications. In Proceedings of the 13th Annual ACM International Conference on Multimedia, Singapore, 6–11 November 2005; pp. 351–354.
13. Levin, A.; Lischinski, D.; Weiss, Y. Colorization Using Optimization. Available online: https://www.researchgate.net/publication/2896183_Colorization_using_Optimization (accessed on 4 March 2023)
14. 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, Republic of Korea, 22–26 October 2018; pp. 1536–1544.
15. Zhang, L.; Li, C.; Wong, T.T.; Ji, Y.; Liu, C. Two-stage sketch colorization. *ACM Trans. Graph. (TOG)* **2018**, *37*, 1–14.
16. 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, Nashville, TN, USA, 20–25 June 2021; pp. 3946–3950.
17. 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, Nashville, TN, USA, 20–25 June 2021; pp. 9889–9898.
18. Sato, K.; Matsui, Y.; Yamasaki, T.; Aizawa, K. Reference-based manga colorization by graph correspondence using quadratic programming. In Proceedings of the SIGGRAPH Asia 2014 Technical Briefs, Shenzhen, China, 3–6 December 2014; pp. 1–4.
19. Huang, J.; Liao, J.; Kwong, S. Semantic example guided image-to-image translation. *IEEE Trans. Multimed.* **2020**, *23*, 1654–1665.
20. 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.
21. 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.
22. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144.
23. Chui, H.; Rangarajan, A. A new algorithm for non-rigid point matching. In Proceedings of the Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662), Hilton Head, SC, USA, 15 June 2000; Volume 2, pp. 44–51.

24. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. In Proceedings of the Medical Imaging with Deep Learning, Amsterdam, The Netherlands, 4–6 July 2018.
25. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* 2015, arXiv:1409.1556.
26. Huber, P.J. Robust estimation of a location parameter. In *Breakthroughs in Statistics: Methodology and Distribution*; Springer: Berlin/Heidelberg, Germany, 1992; pp. 492–518.
27. 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.
28. 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, 18–23 June 2018; pp. 8798–8807.
29. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 694–711.
30. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
31. 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, Republic of Korea, 27–28 October 2019; pp. 9056–9065.
32. Branwen, G.; Gokaslan, A. Danbooru2019: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset. Available online: https://www.gwern.net/Danbooru2019 (accessed on 13 January 2020).
33. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
34. Bynagari, N.B. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Asian J. Appl. Sci. Eng.* **2019**, *8*, 25–34.