Pacific Graphics 2021
E. Eisemann, K. Singh, and F.-L Zhang
(Guest Editors)

Volume 40 (2021), Number 7

Line Art Colorization Based on Explicit Region Segmentation

Ruizhi Cao, Haoran Mo, and Chengying Gao†

Sun Yat-sen University

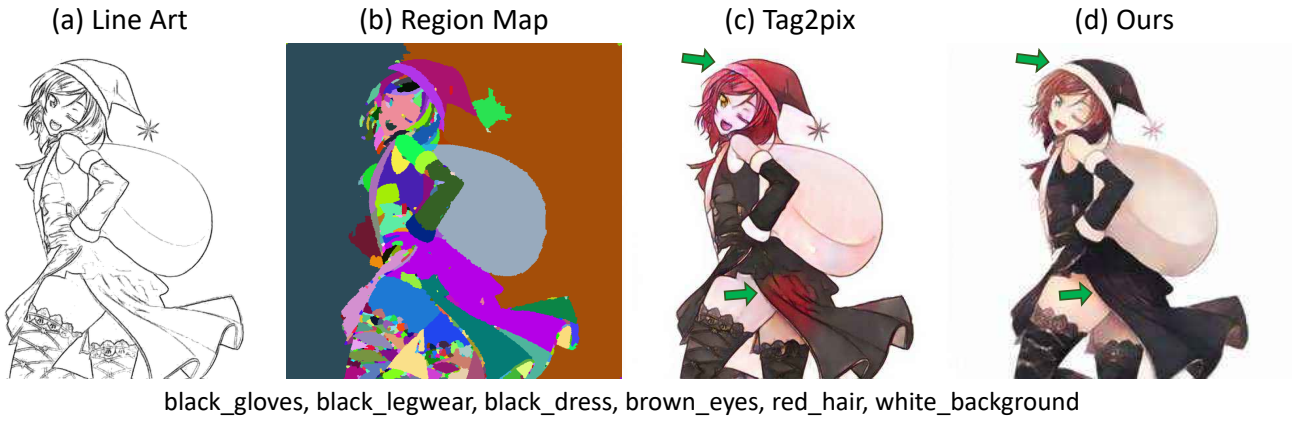

Figure 1: *Comparison of our approach for tag-based line art colorization with a state-of-the-art method. Existing approaches tend to produce bleeding colors. Our proposed explicit segmentation fusion mechanism is able to be incorporated with a variety of line art colorization frameworks, by using region segmentation information explicitly during training to help to alleviate the bleeding artifacts.*

Abstract

Automatic line art colorization plays an important role in anime and comic industry. While existing methods for line art colorization are able to generate plausible colorized results, they tend to suffer from the color bleeding issue. We introduce an explicit segmentation fusion mechanism to aid colorization frameworks in avoiding color bleeding artifacts. This mechanism is able to provide region segmentation information for the colorization process explicitly so that the colorization model can learn to avoid assigning the same color across regions with different semantics or inconsistent colors inside an individual region. The proposed mechanism is designed in a plug-and-play manner, so it can be applied to a diversity of line art colorization frameworks with various kinds of user guidances. We evaluate this mechanism in tag-based and reference-based line art colorization tasks by incorporating it into the state-of-the-art models. Comparisons with these existing models corroborate the effectiveness of our method which largely alleviates the color bleeding artifacts. The code is available at `https://github.com/Ricardo-L-C/ColorizationWithRegion`.

CCS Concepts

• ***Computing methodologies*** → ***Image manipulation;*** *Neural networks; Computer vision;*

1. Introduction

Automatic colorization for line arts plays a critical role in the cartoon and anime industry, because it can reduce a large amount of workload for the professional artists, and thus save a lot of cost for the companies. Given that the commercial value of automatic line art colorization is significantly huge, a number of academic works [QWH06, LQWL18, ZLSS*21] and commercial applications [Yon17, ZLW*18] on this topic have been proposed in recent years.

For all these works or applications, the common goal is to obtain high-quality colorized results that are visually pleasing to hu-

† Corresponding author: mcsgcy@mail.sysu.edu.cn

mans. However, most of them tend to suffer from color bleeding issue. Figure 1-(c) shows a representative result with bleeding colors, which is produced by a state-of-the-art tag-based line art colorization method Tag2pix [KJPY19]. Around the region pointed by the top arrow, the color of the hair (*i.e.,* red) is spread to the brim of the hat, which indicates the same color is assigned across regions with different semantics. The bottom arrow points to an undesired red piece on the dress, indicating inconsistent colors are produced inside an individual region with the only semantic concept. These color bleeding artifacts occur commonly in existing approaches, probably due to the lack of region identification prior to colorization. We propose an explicit segmentation fusion mechanism to incorporate region segmentation information into the colorization method. The segmentation information is used as additional guidance when training the colorization model, so that the model is able to learn to avoid assigning cross-region colors and guarantee the color consistency inside a region. As shown in Figure 1-(d), the bleeding colors are largely reduced.

Currently there are various kinds of user guidances which allow precise line art colorization, *e.g.,* scribbles [SLF*17, ZLW*18], reference images [ZJLL17, LKL*20], textual tags [KJPY19], and language [ZMG*19]. We aim at designing our explicit segmentation fusion mechanism in a plug-and-play manner so that it can be applied to a variety of line art colorization frameworks. We achieve this by introducing two types of fusion modes, both of which can be embedded into a given colorization network while maintaining the overall architecture of the network. An auxiliary loss function is derived from the proposed fusion modes and serves as a complement of the total loss, which helps to improve the potential ability of region segmentation explicitly and alleviate color bleeding simultaneously during the end-to-end training.

We evaluate our proposed mechanism through comprehensive comparisons with existing approaches on tag-based and reference image-based line art colorization tasks. These experiments demonstrate that our proposed explicit segmentation fusion mechanism works well with different kinds of colorization frameworks, and helps to overcome the color bleeding issue.

The main contributions of this work are summarized as follows:

- An explicit segmentation fusion mechanism that aids line art colorization process in alleviating color bleeding artifacts.
- Two types of plug-and-play fusion modes that allow the proposed mechanism to be applicable to a variety of colorization frameworks with different kinds of user guidances.
- In-depth comparisons with existing methods in different user-guided line art colorization tasks.

2. Related Work

2.1. Generative Adversarial Networks

Due to the high-quality performance, generative adversarial networks (GANs) [GPAM*14] are now widely used in image-to-image translation tasks such as image super-resolution [LTH*17, WYW*18], image denoising [CCCY18, YYZ*18], image colorization [LYP*20, VRB20, KJPY19], etc. Currently, there exist a number of works that incorporate additional information into the original GAN model to better control the image generation process, *e.g.,* input image as condition during training [IZZE17], or category label as side information [OOS17, CH18]. In this work, we adopt explicit region segmentation information as additional guidance for the colorization process.

2.2. Line Art Colorization

Automatic colorization approaches [LMS16, ZIE16, ISSI16] generate diverse images with reasonable colors in human perception. While for some special kinds of images, *e.g.,* line art, user guidance as an additional input is necessary to allow for accurate control of the colorization process. Several types of user guidances have been adopted.

Some methods adopt scribbles, *i.e.,* color strokes or dots, as user guidance to provide spatially accurate color instructions. AlacGAN [CMW*18] inserts the scribble map into down-sampling layers of the encoder in the generator whose size is $1/16$ of the input line art. This decreases the accuracy of the scribble information and causes color bleeding. Style2Paints-V3 [ZLW*18] goes further by proposing a two-stage method that allows users to fix incorrect or bleeding colors. It tackles the color bleeding issue in a semi-automatic way with user interaction. In contrast, our approach overcomes this issue automatically.

Reference image is also used as a kind of user guidance. Style2Paints-V1 [ZJLL17] and SCFT [LKL*20] design independent encoders to process sketch and reference image, respectively. The two methods transfer color information from the reference image to the sketch so that the colorized sketch shares similar color content with the reference. Cross domain images translation networks such as MUNIT [HLBK18] and CoCosNet [ZZC*20] can also be employed to colorize line art with reference image. In all these approaches, incorrect color attribution tends to occur when the reference image has different image contents from the ones of the source image, which leads to color bleeding and color mixing readily.

Language or textual tags are also adopted as a natural and convenient user input as additional guidance. SketchSceneColorization [ZMG*19] uses several long sentences describing different objects and different colors to colorize scene sketches [ZYD*18] step by step. Tag2pix [KJPY19] uses two categories of textual tags as user instructions for line art colorization, and it is required to learn segmentation implicitly to distinguish the target regions indicated in the tags. In general, this implicit learning is not effective enough and is not able to overcome color bleeding issue well. Language or text provides spatially loose correspondence between color instruction and target regions. Therefore, such models have difficulty in identifying the precise regions and produce results with artifacts.

To overcome the issue of color bleeding, we propose an explicit segmentation fusion mechanism, which enables colorization models to distinguish each individual region and assign colors correctly and consistently.

2.3. Line Art Segmentation

Region-level segmentation is the main component in our approach. There are few works that focus on this direction. Sketch-

Parse [SDBM17] is a data-driven region-based semantic segmentation framework for sketches. It mainly works with simple sketches, and it is difficult to adapt to complicated line drawings, *e.g.,* illustrations, because it is tedious to manually create regional semantic annotation for them.

DanbooRegion [ZJL20] aims at extracting regions from illustrations and cartoon images. The DanbooRegion dataset provides paired illustrations and region maps. The region maps can be converted to skeleton maps which store region segmentation information without semantic meanings. The region maps are translated to skeleton maps because skeleton maps are learnable and can be directly predicted by a neural network given a line art while the region maps are unlearnable (the translation between the two maps is introduced in the supplemental materials). Skeleton maps can be a kind of auxiliary guidance to provide colorization networks with segmentation information, so we adopt it in our approach to alleviate the color bleeding issue. SplitFilling [ZLSS*21] uses skeleton maps in its workflow to improve the quality of the results. The skeleton maps are used as post-processing in SplitFilling, whereas in our approach they are utilized as additional inputs or targets during training.

3. Method

3.1. Overview

Our framework is based on a generative adversarial network (GAN) consisting of a generator G and a discriminator D. The generator takes a grey-scale line art image $x \in \mathbb{R}^{w \times h \times 1}$ and user guidance (*e.g.,* color tags, reference image, etc.) as inputs, and outputs a colorized illustration $\hat{y} \in \mathbb{R}^{w \times h \times 3}$. The discriminator takes the output image or the target image as input, and determines if the input image is real or not.

We aim to reduce the color bleeding artifacts in the colorization results. To overcome this issue, we propose an explicit segmentation fusion mechanism, which instructs the colorization model not to assign cross-region colors by incorporating regional segmentation information. As illustrated in Figure 2, we introduce two types of fusion modes for this mechanism, which are named "**Direct Concatenation**" and "**Dual-branch**", respectively. "Direct Concatenation" model employs an additional U-Net [RFB15] type network to first generate a skeleton map $\hat{s} \in \mathbb{R}^{w \times h \times 1}$ from an input line art image. The predicted skeleton map is subsequently used as an input to the generator of the colorization network as auxiliary guidance. "Dual-branch" model, however, employs an additional branch in the decoding phase of the generator to produce regional segmentation information stored by a skeleton map $\hat{s} \in \mathbb{R}^{w \times h \times 1}$. The intermediate features of this segmentation branch are fused into the original colorization branch in the generator, which enables the joint modeling of the color features and the segmentation features. A loss function derived from the segmentation branch helps to produce a better skeleton map and provide more accurate segmentation features for the colorization process to avoid the color bleeding.

We design two plug-and-play fusion modes, so that the proposed explicit segmentation fusion mechanism is applicable to a wide variety of line art colorization frameworks with different model architectures and different kinds of user guidances, such as scribbles, reference images, textual tags, language, etc. The "Direct Concatenation" approach can be applied to any colorization framework theoretically, because only one more channel is required to place the predicted skeleton map as an additional input. In contrast, "Dual-branch" is limited to U-Net type networks and is not practical to frameworks with complex workflows. For example, we evaluate our proposed mechanism with a reference-based line art colorization task, and apply the mechanism to MUNIT framework [HLBK18] for reference-based image translation, where two auto-encoders are used to reconstruct both images and latent codes from different domains. This framework is not a conventional U-Net model and adding dual branches is likely to destroy the workflow. Therefore, only "Direct Concatenation" is applied to MUNIT. We additionally evaluate for tag-based line art colorization task using Tag2pix [KJPY19] framework. It is a U-Net type network, so both two fusion modes are employed. Therefore, in the following sections, we take tag-based colorization for example to introduce the explicit segmentation fusion mechanism along with the two fusion modes in detail.

3.2. Explicit Segmentation Fusion Mechanism

3.2.1. Direct Concatenation

Given that we are able to use the trained model from DanbooRegion [ZJL20] to extract the region maps from line art images, it is straightforward to input such regional information to the network directly to serve as extra guidance for colorization. As shown in Figure 2-(a), we first use the U-Net type network with the trained weights to produce a skeleton map $\hat{s}$ from the input line art, and then concatenate the line art image and the skeleton map directly. We then input the hybrid image to the generator of the colorization model while keeping the remaining architecture of the generator unchanged.

3.2.2. Dual-branch

The fusion with a dual-branch generator network is designed for learning colorization and regional segmentation simultaneously, and improving the colorization performance by fusing the features of the two processes. As shown in Figure 2-(b), the two branches share a common encoder, which is built on a series of encoding blocks containing convolution layers to extract feature maps from the input line art. In the decoding phase, one branch works on producing the colorized output $\hat{y}$, and the other branch generates regional segmentation information represented in a skeleton map $\hat{s}$. They are both built on stacked decoding blocks, which are skip connected to the mirrored encoding blocks to obtain the hierarchical encoded information directly.

In order to improve the colorization performance and alleviate the color bleeding artifact, we fuse segmentation information into the colorization branch so that the colorization branch can learn to assign colors to each individual region and avoid filling colors across regions with different semantic contents. We propagate the segmentation information by adding connections from the segmentation branch to the colorization branch. Specifically, each decoding block in the colorization branch receives feature maps from

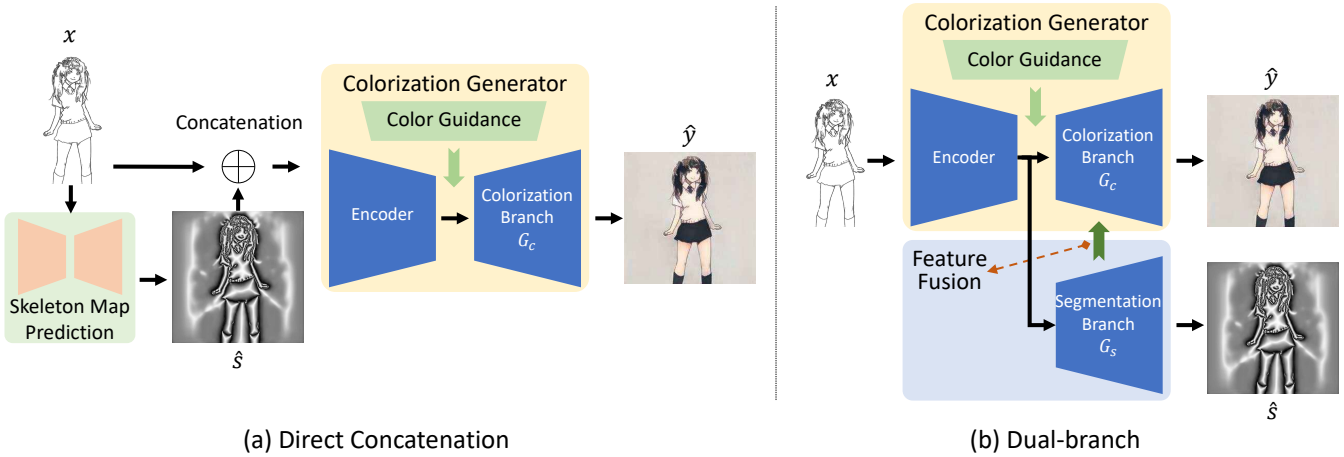

Figure 2: *Network architectures with two types of fusion modes for the explicit segmentation mechanism.*

three sources as inputs: features from the previous decoding block, features from the mirrored encoding block, and output features from the corresponding decoding block in the segmentation branch. These features are concatenated to form hybrid features, which are then processed by the next decoding block.

3.3. Loss Function

We propose two types of fusion modes for the explicit segmentation information, which are applicable to a diversity of line art colorization models. In general, fusion with direct concatenation does not change the architecture of the colorization network much, and therefore the loss functions during training can be the same as the original ones. For the model with a "Dual-branch" fusion mode, an additional branch is introduced, from which an additional loss for penalizing the segmentation performance is derived. In this section, we use a representative tag-based line art colorization approach Tag2pix [KJPY19] as an example to discuss how the new loss for the explicit segmentation is added.

Tag2pix [KJPY19] proposes a two-step training method, aiming at using changing loss to switch the focus of the network between the segmentation step and the colorization step. However, Tag2pix does not utilize explicit segmentation information to guide the training procedure. In contrast, our approach uses explicit segmentation information stored in skeleton maps as guidance for the colorization process. Therefore, we follow the training protocol of Tag2pix with changing loss but add the penalization of the explicit region segmentation.

Segmentation. We expect the segmentation branch to provide segmentation information for the colorization branch, so the segmentation branch should learn to generate plausible segmentation features itself. In the experiments, the training of the segmentation branch seems to converge easily. Therefore, we train it with pixel-wise L_1 distance between the generated skeleton map and ground truth, which is defined as:

$$L_{seg} = \mathbb{E}_{x,s}[||s - G_s(x)||_1], \tag{1}$$

where x is the input line art, s the ground truth skeleton map and $G_s(x)$ the predicted skeleton map from the segmentation branch G_s in the generator. $\mathbb{E}$ denotes expectation and $||\cdot||_1$ is L_1 normalization.

At segmentation step, colorization branch and guide decoder use reconstruction loss and adversarial loss, which are formulated as follows:

$$L_{rec} = \mathbb{E}_{x,y}[||y - G_c(x,t)||_1 + \beta||y - G_g(x,t)||_1], \tag{2}$$

$$L_{adv} = \mathbb{E}_y[\log D_{adv}(y)] + \mathbb{E}_x[\log(1 - D_{adv}(G_c(x,t))], \tag{3}$$

where t denotes the color tags represented in a one-hot vector, y the ground truth color illustration, $G_c(\cdot,\cdot)$ the colorized output from the colorization branch, $G_g(\cdot,\cdot)$ the output from the guide decoder. β in Eq.(2) is a scalar. $D_{adv}(\cdot)$ is the output of discriminator that judges whether the generated color image is real or fake.

The total losses for discriminator and generator at segmentation step are respectively defined as:

$$L_D = -L_{adv}, \tag{4}$$

$$L_G = L_{adv} + \lambda_{rec}L_{rec} + \lambda_{seg}L_{seg}, \tag{5}$$

where λ_{rec} and λ_{seg} are scalars.

Colorization. At colorization step, tag classification loss L_{cls} is added. It serves as an auxiliary classifier, which gives the discriminator instructions on distinguishing color tags:

$$L_{cls} = \mathbb{E}_{y,t}[-\log D_{cls}(t|y)] + \mathbb{E}_{x,t}[-\log D_{cls}(t|G_c(x,t))], \tag{6}$$

where $D_{cls}(t|y)$ denotes binary classification for each color tag t, given y.

The total losses for discriminator and generator at colorization step are respectively defined as:

$$L_D = -L_{adv} + \lambda_{cls}L_{cls}, \tag{7}$$

$$L_G = L_{adv} + \lambda_{cls}L_{cls} + \lambda_{rec}L_{rec} + \lambda_{seg}L_{seg}, \tag{8}$$

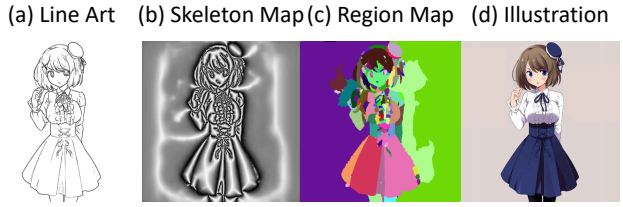

Figure 3: *An example in our dataset. (a) Line art extracted by SketchKeras [lll17]. (b) Skeleton map predicted by DanbooRegion [ZJL20]. (c) Region map converted from the skeleton map through the watershed algorithm [NP14]. (d) Ground truth illustration from the Danbooru dataset [AcB21].*

where λ_{cls}, λ_{rec} and λ_{seg} are scalars.

4. Experiments

4.1. Dataset

Our proposed explicit segmentation fusion mechanism is able to be applied to different kinds of line art colorization frameworks. We evaluate this mechanism on a tag-based and a reference image-based line art colorization task. Both tasks take a line art image and their corresponding guidance as inputs and generates a colorized output. We use the illustrations in Danbooru dataset [AcB21] as the ground truth for the colorization branch. The input line arts can be extracted from the color illustrations through existing techniques [lll17, MSSG*21, SSII18] and we found SketchKeras [lll17] works best in practice. For the tag-based colorization task, we follow how Tag2pix [KJPY19] did to select the tags from the Danbooru dataset. During this process, however, we observed there are duplicated or similar images in the Danbooru dataset. In order to remove the redundant data and make the training more efficient, we calculated the Perceptual Similarity [ZIE*18] between two images with similar tags, and discarded one of them if the similarity score is high. For the reference-based task, we follow the protocol in MUNIT [HLBK18] and select a color illustration randomly as the reference image.

Ground truth skeleton maps are required for the segmentation branch in the "Dual-branch" fusion mode. We notice a recent dataset DanbooRegion [ZJL20], which is built on the Danbooru dataset with region annotations for illustrations. However, it lacks the tag information and loses the identity mapping to the original dataset, and thus is not practical in our experiments. Given that this work provides a trained model for region extraction, which seems to work well with both illustrations and line arts, we utilize this model to produce skeleton maps from the line art images and use them as the ground truth. An example of the quadruple data is shown in Figure 3.

4.2. Implementation Details

In the tag-based colorization task, we apply our proposed mechanism to Tag2pix [KJPY19] with a two-step training. We train 5 epochs for the segmentation step first, and 35 epochs for the colorization step subsequently. Learning rates for the generator and the discriminator are both set to 0.0002. In Eq.(5) and Eq.(8), we use $\lambda_{rec} = 1000$ and $\lambda_{seg} = 0.9$. In Eq.(2), we set $\beta = 0.9$. λ_{cls} in Eq.(7) and Eq.(8) are both set to 40 for CITs and CVTs. During testing, we use both extracted and real line arts to evaluate the colorization performance. For each line art, the color tags are manually assigned.

Figure 4: *Effectiveness of alleviating color bleeding compared with Tag2pix [KJPY19]. Our results are from the framework with a dual branch.*

As for the reference-based colorization task, MUNIT [HLBK18] is used as our baseline method and we apply our mechanism to it. We use the same hyperparameters as the official implementation of MUNIT. Specifically, learning rates for the generator and the discriminator are set to 0.0001 and 0.0004. We train the network for 140,000 iterations with an adversarial loss, an image reconstruction loss, a content reconstruction loss, a style reconstruction loss and a cycle reconstruction loss.

4.3. Effectiveness of Avoiding Color Bleeding

Our proposed approach with an explicit region segmentation fusion mechanism is designed for overcoming color bleeding issues. We first evaluate the performance of this mechanism on the tag-based line art colorization task and compare it with Tag2pix [KJPY19] that learns implicit regional segmentation while achieving colorization.

Qualitative Results. The comparisons are shown in Figure 4. Tag2pix often produces bleeding colors inside a region with a common semantic concept. For example, in the first and second rows, it produces different colors and color gradients in the arms. In contrast, our approach draws the only color smoothly for the arms, which makes the colorized output visually more pleasing. The region maps show that the segmentation branch in our framework learns to identify the whole arm in the line arts from the first and second rows as a complete region. With this additional information as guidance, it is easier for the colorization branch to learn to assign the arm with a consistent color.

We further evaluate the performance on overcoming color bleeding issue by inputting multiple tags. As shown in Figure 5, Tag2pix [KJPY19] generates hybrid colors in the hemline of the dress in the top example, and in the inner region of the dress in the bottom example, when given different tags as colorization instructions. From our results, we can see that such artifacts are largely alleviated and the colors are smooth and consistent inside the corresponding regions for whatever input color tags. These results further demonstrate that the improvement with less bleeding colors during the colorization process is not caused by factors such as the input tags or the randomness during training, but by the segmentation information we explicitly inject into the colorization process. It is also proved that utilizing explicit regional segmentation as guidance during training is superior to the implicit learning of regional segmentation in Tag2pix [KJPY19]. Please refer to the supplemental materials for more qualitative results.

Quantitative Evaluation and User Study. We use Fréchet Inception Distance (FID) [HRU*17], which serves as a metric to measure the similarity between two sets of images, to quantitatively evaluate the overall quality of the generated results. We randomly select 500 illustrations from the Danbooru dataset [AcB21], and extract the corresponding line drawings and tags as the test set. Then, for each method, we generate the color results with the line drawings and the tags, and calculate the FID scores with these results and the illustrations. The results are shown in Table 1. We can see that our approach has a lower FID score, which indicates better overall quality of our results in comparison with those from Tag2pix [KJPY19].

Given that FID only measures the overall quality and cannot correctly reflect the performance on avoiding the color bleeding artifacts, we conduct an additional user study to evaluate the effectiveness of overcoming this issue of our approach with two fusion modes. We follow the evaluation criteria in the user study in Tag2pix, and measures three aspects: (1) *Less Color Bleeding*: how well the model produces less color bleeding; (2) *Tag Accuracy*: how well the colorized results match the instructive tags; (3) *Overall Quality*: how high the overall quality of the colorized results is. We recruited 22 participants for this study. 40 examples were selected randomly from the test set to form the study samples. For each sample generated by Tag2pix and our approach, they ranked 1 (worst) to 5 (best) for each of the three aspects. The ranking scores of all the participants are averaged finally and shown in Table 2. In the aspect of *Less Color Bleeding*, both two modes of our method have better scores, which confirms that our proposed explicit segmentation fusion mechanism does take effect on overcoming this issue. The scores of *Tag Accuracy* and *Overall Quality* both show that the incorporation of segmentation information improves the faithfulness to the input tags and the overall colorization performance.

Table 1: *Quantitative comparisons between Tag2pix [KJPY19] and our approach on FID score. A lower value is better.*

Method	FID (↓)
Tag2pix	76.744
Ours (Dual-branch)	**64.482**

Table 2: *User study of the performance of Tag2pix [KJPY19] and our approach. 'Concat' denotes Direct Concatenation and 'Dual' Dual-branch. The increments in the brackets are based from Tag2pix. For all the three aspects, a higher value is better.*

Categories	Tag2pix	Ours (Concat)	Ours (Dual)
Less Color Bleeding	1.66	2.95 (+1.29)	**4.25** (+2.59)
Tag Accuracy	1.98	3.25 (+1.27)	**4.16** (+2.18)
Overall Quality	1.84	3.19 (+1.35)	**4.34** (+2.50)

4.4. In-depth Study

4.4.1. Ablation Study on Fusion Mode

We evaluate the two types of fusion modes, "Direct Concatenation" and "Dual-branch", in our proposed explicit segmentation information fusion mechanism qualitatively and quantitatively with the tag-based line art colorization task. The qualitative results are shown in Figure 6. Compared with results from Tag2pix [KJPY19] that suffer from the color bleeding issue, the two approaches both avoid such an artifact to some extent. Fusing segmentation information with a dual branch performs better than the method with direct concatenation in most cases of the test set, especially on assigning consistent colors to semantically symmetric regions and matching the regions with the tags. For instance, in the examples from the first and second rows, the "Direct Concatenation" model draws the left and right parts of the dress and the collars with slightly different colors although the two parts are semantically symmetric. In contrast, the "Dual-branch" model is more likely to produce similar colors. In the bottom row example, "Direct Concatenation" model paints a part of the hair in red, which is the color from the "eyes", while the "Dual-branch" model is able to assign a consistent color for the hair.

Quantitative results in Table 2 also indicate that users prefer results from the "Dual-branch" mode, which are in line with the visual results. We believe this is because in "Direct Concatenation" mode, part of information of the skeleton map might be lost due to the down-sampling process in the encoder prior to the colorization stage. While with "Dual-branch", the features of the skeleton map are directly fused with the colorization branch in a pyramid manner, which provides enough segmentation information and thus results in better performance. Although "Dual-branch" mode is superior to "Direct Concatenation" on the whole, the latter does produce better

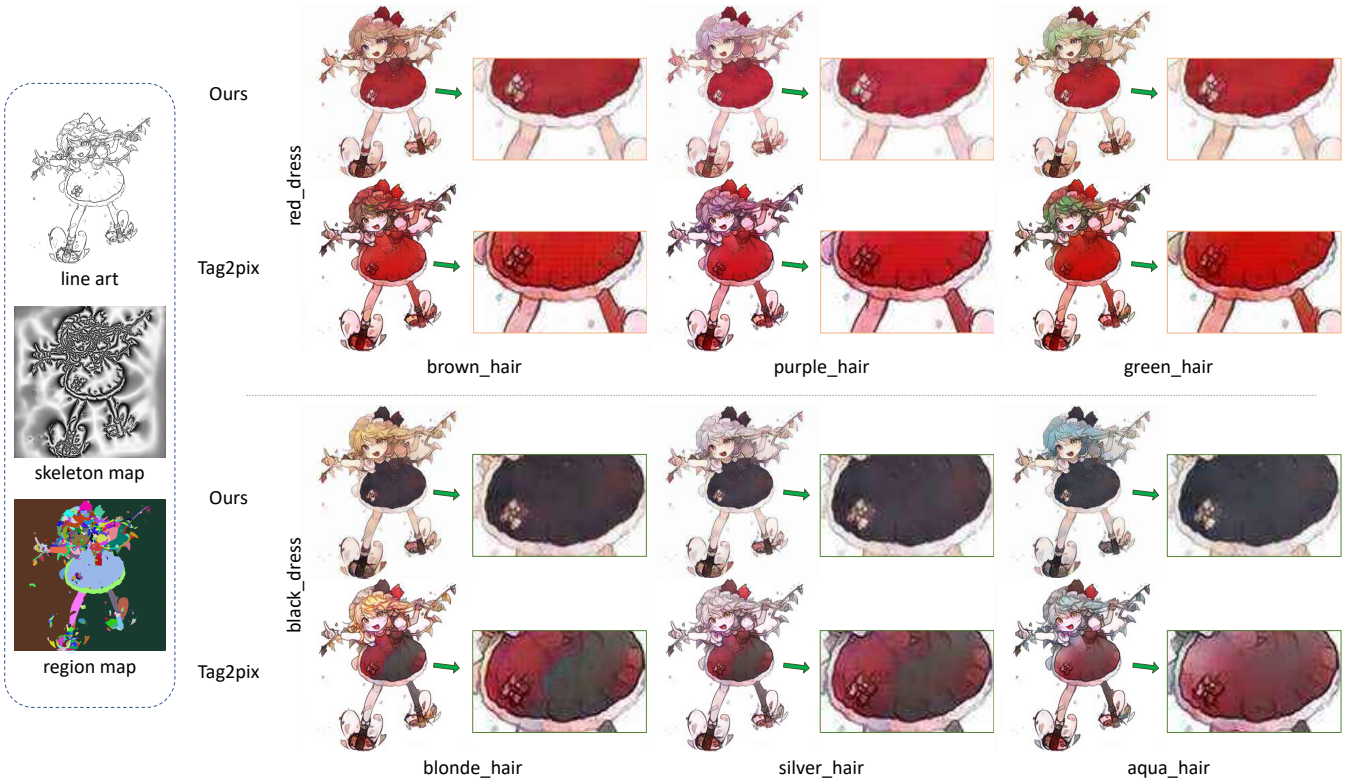

Figure 5: *Colorization with various tags. "white_shirt", "red_eye", "red_footware", and "white_background" are common tags for all examples. Our results are from the framework with a dual branch.*

Figure 6: *Effectiveness of the two fusion modes of the segmentation information.*

results in a fraction of examples. We show these results and more examples of the comparisons in the supplemental materials.

4.4.2. Improvement on Color Contrast

We also notice some improvement brought by adding the explicit segmentation information. As can be seen in Figure 7, Tag2pix [KJPY19] may fill a number of regions with similar colors in some cases, while our approach seems to generate images with higher color contrast, that is, richer colors. We guess this is because the hybrid learning of colorization and implicit regional segmentation in Tag2pix makes it hard to learn correct color assignment. As a result, drawing with average colors in different regions could be a safer strategy during training. On the contrary, when adding segmentation information for explicit guidance, our framework learns in a easier direction and thus is able to capture richer colors in the target color illustrations. We show more results of the comparison in the supplemental materials.

4.5. Reference-based Colorization

The idea of adding segmentation information into the colorization procedure as explicit hints to avoid color bleeding can be applied to line art colorization frameworks with various kinds of user guidances. Therefore, we additionally evaluate the proposed idea on a reference image-based line art colorization task. This task can be

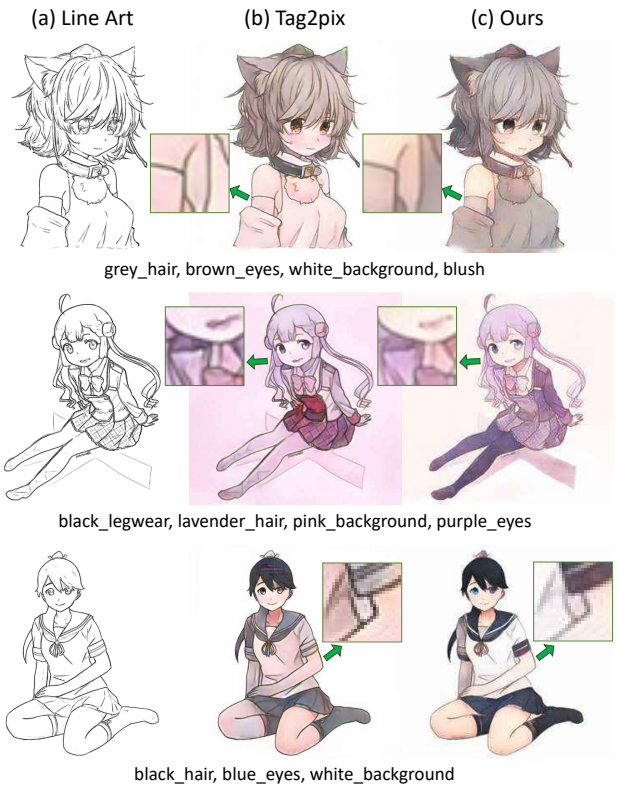

Figure 7: *Comparisons on color contrast between different methods. Our results are from the framework with a dual branch.*

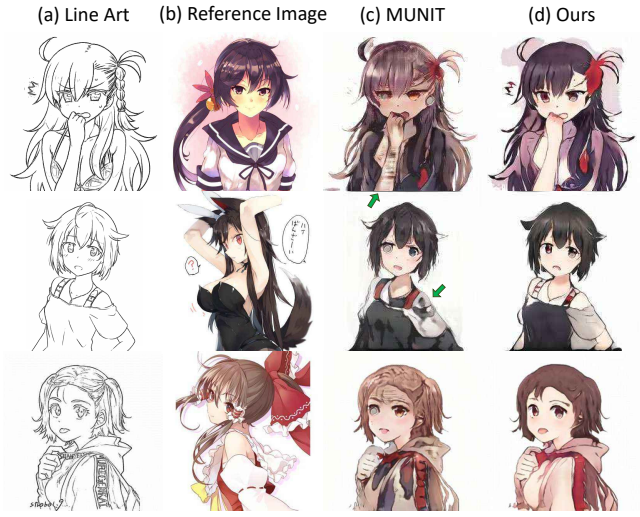

Figure 8: *Results on reference-based line art colorization. Our results are from model by incorporating MUNIT [HLBK18] with explicit segmentation information in a direct concatenation mode.*

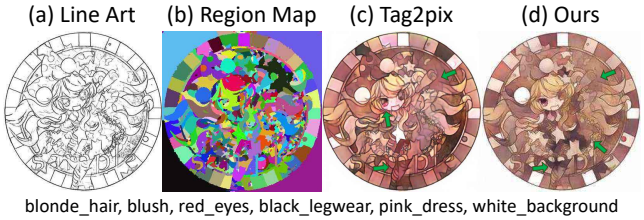

Figure 9: *Limitation of our approach on a highly complicated example with dense strokes. The predicted regions are fragmentary and fail to help to improve the colorization performance. Bleeding colors and regions that do not match the tags are pointed by the arrows. Our result is from the model with a dual-branch.*

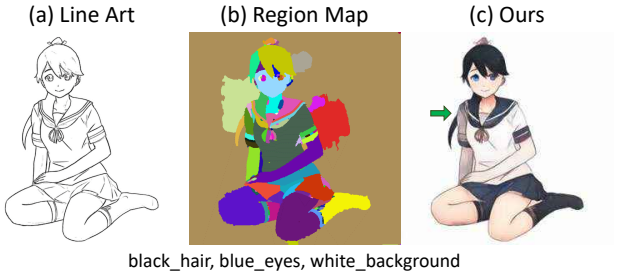

Figure 10: *Limitation of our approach on assigning colors for regions with the same semantic concepts. Our result is from the model with a dual-branch.*

regarded as an unsupervised style transfer problem, so we employ a representative method MUNIT [HLBK18] as the baseline method. While MUNIT has a complicated network architecture, we fuse the segmentation information into this model through direct concatenation instead of a dual-branch.

The qualitative comparisons between MUNIT and our method are shown in Figure 8. MUNIT tends to generate bleeding colors (*e.g.,* the black shading on the arm in top row example), which causes a worse overall quality. When incorporated with our proposed segmentation information fusion mechanism, most individual parts are assigned with consistent colors in the results. This is because even with a straightforward fusion of the explicit segmentation information, the colorization model is able to obtain additional hints on how to assign colors correctly and thus learns to generate results with higher quality. These results confirm that our proposed explicit segmentation fusion mechanism is able to work with different kinds of line art colorization frameworks and does help to improve the colorization performance by alleviating the bleeding artifacts. Please refer to the supplemental materials for more results of the comparison.

5. Limitations and Discussion

Our approach that explicitly fuses segmentation information with the colorization process learns to assign consistent and smooth colors to the individual regions, and thus is able to alleviate the color

bleeding issue. However, it may fail on highly complicated examples with dense lines. Figure 9 shows an example in this case, in which the predicted regions are fragmentary and small in size due to the dense strokes in the line art image. Such a dense region map causes redundant segmentation information, and is difficult for the model to connect the regions with the same semantic content (*e.g.,* the hair) and guarantee the color consistency in these regions. Therefore, our approach works less than satisfactory on overcoming the color bleeding issue, and has similar performance to Tag2pix [KJPY19]. This problem could be addressed by incorporating a two-stage type method [ZLW*18] with additional user interaction to connect the small regions with similar semantics and refine the results, which may be a promising future direction.

The output regional segmentation maps in our approach help to fill colors that are more faithful to the tags and more consistent for semantically symmetric regions in some cases, as discussed in Section 4.4.1. Nevertheless, the predicted segmentation maps do not contain the semantic concepts of the regions essentially. As a result, in some cases where two adjacent regions have the same semantic concepts, our model might still fail to connect these two regions and paint them with different colors. A representative example is shown in Figure 10. The sleeve pointed by the arrow is a part of the clothes, but it belongs to an individual region adjacent to the main body of the clothes. The colors of the two parts are significantly different, leading to low visual quality. To enable the model to understand the semantic content of the segmented regions from anime characters, transfer learning on the human parsing [ZLC*18] or face parsing [LLWL20] datasets could be incorporated to form a further extension of our work.

Acknowledgments

This work was supported by the Natural Science Foundation of Guangdong Province, China (Grant No. 2019A1515011075) and the National Science Foundation of China under Grant U1811262, Grant 61772567.

References

[AcB21] ANONYMOUS, COMMUNITY D., BRANWEN G.: Danbooru2020: A large-scale crowdsourced and tagged anime illustration dataset. https://www.gwern.net/Danbooru2020, January 2021. Accessed: DATE. URL: https://www.gwern.net/Danbooru2020.

[CCCY18] CHEN J., CHEN J., CHAO H., YANG M.: Image blind denoising with generative adversarial network based noise modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3155–3164.

[CH18] CHEN W., HAYS J.: Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 9416–9425.

[CMW*18] 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* (2018), pp. 1536–1544.

[GPAM*14] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. In *Advances in Neural Information Processing Systems* (2014), Ghahramani Z., Welling M., Cortes C., Lawrence N., Weinberger K. Q., (Eds.), vol. 27, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[HLBK18] 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)* (2018), pp. 172–189.

[HRU*17] HEUSEL M., RAMSAUER H., UNTERTHINER T., NESSLER B., HOCHREITER S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS* (2017).

[ISSI16] IIZUKA S., SIMO-SERRA E., ISHIKAWA H.: Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG) 35*, 4 (2016), 1–11.

[IZZE17] ISOLA P., ZHU J.-Y., ZHOU T., EFROS A. A.: Image-to-image translation with conditional adversarial networks. *CVPR* (2017).

[KJPY19] 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* (2019), pp. 9056–9065.

[LKL*20] 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* (2020), pp. 5801–5810.

[lll17] LLLYASVIEL: sketchkeras. URL: https://github.com/lllyasviel/sketchKeras.

[LLWL20] LEE C.-H., LIU Z., WU L., LUO P.: Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).

[LMS16] LARSSON G., MAIRE M., SHAKHNAROVICH G.: Learning representations for automatic colorization. In *European conference on computer vision* (2016), Springer, pp. 577–593.

[LQWL18] LIU Y., QIN Z., WAN T., LUO Z.: Auto-painter: Cartoon image generation from sketch by using conditional wasserstein generative adversarial networks. *Neurocomputing 311* (2018), 78–87.

[LTH*17] LEDIG C., THEIS L., HUSZÁR F., CABALLERO J., CUNNINGHAM A., ACOSTA A., AITKEN A., TEJANI A., TOTZ J., WANG Z., ET AL.: Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 4681–4690.

[LYP*20] LU P., YU J., PENG X., ZHAO Z., WANG X.: Gray2colornet: Transfer more colors from reference image. In *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 3210–3218.

[MSSG*21] MO H., SIMO-SERRA E., GAO C., ZOU C., WANG R.: General virtual sketching framework for vector line art. *ACM Transactions on Graphics (TOG) 40*, 4 (2021), 1–14.

[NP14] NEUBERT P., PROTZEL P.: Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms. In *2014 22nd international conference on pattern recognition* (2014), IEEE, pp. 996–1001.

[OOS17] ODENA A., OLAH C., SHLENS J.: Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning* (2017), PMLR, pp. 2642–2651.

[QWH06] QU Y., WONG T.-T., HENG P.-A.: Manga colorization. *ACM Transactions on Graphics (TOG) 25*, 3 (2006), 1214–1220.

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (2015), Springer, pp. 234–241.

[SDBM17] SARVADEVABHATLA R. K., DWIVEDI I., BISWAS A., MANOCHA S.: Sketchparse: Towards rich descriptions for poorly drawn sketches using multi-task hierarchical deep networks. In *Proceedings of the 25th ACM international conference on Multimedia* (2017), pp. 10–18.

[SLF*17] SANGKLOY P., LU J., FANG C., YU F., HAYS J.: Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 5400–5409.

[SSII18] SIMO-SERRA E., IIZUKA S., ISHIKAWA H.: Mastering sketching: adversarial augmentation for structured prediction. *ACM Transactions on Graphics (TOG) 37*, 1 (2018), 1–13.

[VRB20] VITORIA P., RAAD L., BALLESTER C.: Chromagan: adversarial picture colorization with semantic class distribution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2020), pp. 2445–2454.

[WYW*18] WANG X., YU K., WU S., GU J., LIU Y., DONG C., QIAO Y., CHANGE LOY C.: Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (2018), pp. 0–0.

[Yon17] YONETSUJI T.: Paintschainer. https://github.com/pfnet/PaintsChainer, 2017. URL: https://github.com/pfnet/PaintsChainer.

[YYZ*18] YANG Q., YAN P., ZHANG Y., YU H., SHI Y., MOU X., KALRA M. K., ZHANG Y., SUN L., WANG G.: Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging 37*, 6 (2018), 1348–1357.

[ZIE16] ZHANG R., ISOLA P., EFROS A. A.: Colorful image colorization. In *European conference on computer vision* (2016), Springer, pp. 649–666.

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR* (2018).

[ZJL20] ZHANG L., JI Y., LIU C.: Danbooregion: An illustration region dataset. In *European Conference on Computer Vision (ECCV)* (2020).

[ZJLL17] ZHANG L., JI Y., LIN X., LIU C.: Style transfer for anime sketches with enhanced residual u-net and auxiliary classifier gan. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)* (2017), IEEE, pp. 506–511.

[ZLC*18] ZHAO J., LI J., CHENG Y., SIM T., YAN S., FENG J.: Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In *Proceedings of the 26th ACM international conference on Multimedia* (2018), pp. 792–800.

[ZLSS*21] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021).

[ZLW*18] ZHANG L., LI C., WONG T.-T., JI Y., LIU C.: Two-stage sketch colorization. *ACM Transactions on Graphics (TOG) 37*, 6 (2018), 1–14.

[ZMG*19] ZOU C., MO H., GAO C., DU R., FU H.: Language-based colorization of scene sketches. *ACM Transactions on Graphics (TOG) 38*, 6 (2019), 1–16.

[ZYD*18] ZOU C., YU Q., DU R., MO H., SONG Y.-Z., XIANG T., GAO C., CHEN B., ZHANG H.: Sketchyscene: Richly-annotated scene sketches. In *Proceedings of the european conference on computer vision (ECCV)* (2018), pp. 421–436.

[ZZC*20] ZHANG P., ZHANG B., CHEN D., YUAN L., WEN F.: Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 5143–5153.