Reference-guided structure-aware deep sketch colorization for cartoons

Xueting Liu¹, Wenliang Wu², Chengze Li¹ (✉), Yifan Li², and Huisi Wu²

© The Author(s) 2021.

Abstract Digital cartoon production requires extensive manual labor to colorize sketches with visually pleasant color composition and color shading. During colorization, the artist usually takes an existing cartoon image as color guidance, particularly when colorizing related characters or an animation sequence. Reference-guided colorization is more intuitive than colorization with other hints, such as color points or scribbles, or text-based hints. Unfortunately, reference-guided colorization is challenging since the style of the colorized image should match the style of the reference image in terms of both global color composition and local color shading. In this paper, we propose a novel learning-based framework which colorizes a sketch based on a color style feature extracted from a reference color image. Our framework contains a color style extractor to extract the color feature from a color image, a colorization network to generate multi-scale output images by combining a sketch and a color feature, and a multi-scale discriminator to improve the reality of the output image. Extensive qualitative and quantitative evaluations show that our method outperforms existing methods, providing both superior visual quality and style reference consistency in the task of reference-based colorization.

Keywords sketch colorization; image style editing; deep feature understanding; reference-based image colorization

1 Introduction

With the increasing popularity of digital cartoons, various computer-assisted technologies for producing

them have rapidly been developed in recent years. Colorizing cartoon sketches is such a technology that has attracted extensive research focus in the field of computer graphics. Since the sketch itself contains no hint about the colors to apply, existing methods either colorize the sketch by pure guesswork [1–3], which may lead to unnatural colors (see Fig. 1(b)), or by using user-provided hints, such as color points or scribbles [4, 5], or text hints [6]. However, manually crafting color hints is time-consuming, especially for line drawings with complex content. For example, in Fig. 1(c), 62 color points were created by the user to obtain a relatively satisfactory colorized result. Besides, crafting proper color hints is also challenging, especially for amateur users, as it requires the user to apply a certain level of aesthetic judgement to generate a visually pleasant color cartoon. Moreover, when coloring a cartoon animation, it would be extremely difficult for the user to achieve color consistency across frames when color hints are provided for each frame individually.

To resolve the above issues, we propose an automatic system that can colorize a cartoon line drawing based on a reference image. There are two key advantages in reference-guided colorization over color hint-based methods. Firstly, it saves the effort of user trial-and-error in creating a proper set of user-hints for colorizing the sketches. Instead, the user only needs to provide a reference image in a similar color style to that desired in the colorized image. The system can automatically learn the color style from this reference image and apply it to colorize the sketch. Secondly, when the user needs to colorize a set of sketches with similar content, such as a sequence of sketched frames for producing a cartoon animation or a set of sketched character designs for the same cartoon character in different poses, the user only

¹ Caritas Institute of Higher Education, Hong Kong SAR, China. E-mail: X. Liu, tliu@cihe.edu.hk; C. Li, czli@cihe.edu.hk (✉).

² Shenzhen University, Shenzhen 518060, China. E-mail: W. Wu, wuwenliang2019@email.szu.edu.cn; Y. Li, 1810272030@email.szu.edu.cn; H. Wu, hsu@szu.edu.cn.

Manuscript received: 2021-01-21; accepted: 2021-03-18

Fig. 1 Comparison between our method and existing sketch colorization and style transfer methods.

needs to colorize one of the sketches in the set and can then transfer the color of this image to the other images with ease.

In the current literature, very few works focus on reference-based colorization and most suffer from low colorization quality: it is challenging to properly propagate the visual style of the reference to the sketch input. However, various existing techniques tailored for other applications may be borrowed for this application. Style transfer methods [7, 8] take an input image and a style image as input and transfer the style of the style image to the input. While these methods may be adapted to apply the style of the reference image to the input sketch in our application, the structural lines in the sketch generally cannot be well preserved, and the result may exhibit obvious artifacts (see Fig. 1(e)). Content-style disentanglement methods separate style from content [9, 10], but they generally cannot ensure the content component to correspond exactly to the structural lines in the sketch. As a result, the encoded style space is usually not perpendicular to the content space, which further affects the quality of the colorized result.

In this paper, we propose a novel learning-based solution for colorizing cartoon sketches based on reference images. A key requirement of our application is that structural lines in the sketch should be preserved during colorization. To do so, we formulate the colorization network as an image-to-image translation problem where the input is a

cartoon sketch and the output is a colorized version of it based on the color style of a reference image. To colorize the sketch based on a reference image, we extract the color style of the reference image and then fuse the extracted color style feature into a deep hierarchical representation of the sketch via adaptive instance normalization (AdaIN) [7]. In order to improve the visual quality of the generated color output, we adopt a multi-scale discriminator which can improve the realism of the output image in terms of both global color composition and local color shading. We have applied our method to various images, with convincing results obtained in all cases.

The contributions of our work can be summarized as follows:

- a novel reference-based cartoon sketch colorization method using a deep learning approach which does not need manual color hinting,
- a formulation of sketch colorization as an image-to-image translation problem where the structural lines in the sketch should be faithfully preserved, and
- a multi-scale discriminator to improve the visual realism of the generated cartoon in terms of both global color composition and local color shading.

2 Related work

2.1 Background

With the rapid development of digitization, many ways have been proposed to solve the difficulties in cartoon creation. For example Ref. [11] proposed using geodesic distance to recolor cartoon images, Ref. [12] tried to use 2D sketches to reconstruct the symmetric in 3D free-form shapes, and Ref. [13] used the shading of a cartoon image to estimate the reflectance and shape of objects. Similarly, there are many methods used for sketch colorization. Among them, plenty perform reference-guided colorization. We can roughly divide these methods into three major categories: image style transfer, conditional image-to-image generation, and style-content disentanglement.

2.2 Image style transfer

Image style transfer aims to alter the style of an image, including its color and texture, based on a style exemplar. Ref. [14] proposed a style editing method which learns the texture mapping from pyramidal features based on the exemplar. However,

its need for paired supervision greatly restricts its usage in our colorization task. Recently, thanks to the revamped construction of image features with deep neural network models, style transfer tasks can achieve higher output quality and style editing precision. Among these methods, Ref. [15] used a content loss and a style loss to handle different image components in the feature domain to achieve style transfer. Ref. [16] further used perceptual loss as a universal metric for all style-related editing tasks. Ref. [7] extended the style transfer method to perform on-the-fly style alteration with a single neural network using a feature rescaling technique, adaptive instance normalization. Later, further methods were proposed to achieve better visual quality [8, 17–20], or efficiency [21, 22]. However, style transfer methods are usually designed with global style objectives and cannot be directly applied to propagate local characteristics of the reference style. In other words, these methods cannot apply customized local color and textures according to underlying structural constraints (see Fig. 1(e)). In contrast, our method not only extracts the global color style to guide colorization, but also cares about detailed local color shading.

2.3 Deep conditional image-to-image synthesis

Deep generative models have been widely used to accomplish cross-domain image-to-image translation tasks. These methods apply adversarial training [23] to synthesize images with natural looks. Training can be either conducted under paired supervision [1] or cycle-consistency [2, 24]. While these methods achieved decent image translation quality and can be used for sketch colorization [3], these methods are generally deterministic, i.e., they cannot create diversified outputs based on different user requirements.

Conditional image generation methods extended the generative model to allow conditional user inputs. Refs. [4, 25] proposed use of scribbled or point color hints as the conditional input to guide colorization. However, they cannot use full-page cartoon pictures as a reference. Additionally, placing color hints is time-consuming and requires users' experiences. Ref. [6] proposed reading a list of text-level visual tags to guide decoration of the sketch input with the given visual properties. However, these visual tags may not well capture the color style the user

targets, unlike a reference images. Ref. [5] attempted to apply the style of the reference cartoon image to an arbitrary sketch input with a feature-based encoder and decoder design. However, the output quality is poor, with blurring artifacts. Furthermore, the color style of the result may be inconsistent with that of the reference image. Unlike the above conditional colorization methods that need explicit visual cues, Ref. [26] proposed encoding visual styles into a latent space with a style encoder. During the image translation process, the style of the reference image is extracted as the conditional input. However, it lacks a strong style editing mechanism for complex datasets such as comics and cartoons, which contain almost unlimited color and texture combinations. More recently, Ref. [27] proposed a reference-based sketch colorization method based on a transformation-aware attention module. However, this method still shares the problem of color style inconsistency between the reference and the output. We will further demonstrate this weakness in Section 3.2.4.

2.4 Reference-based photo colorization

Many works attempt to colorize photographs with reference-based priors. The pioneering work [28] proposed to transfer chromatic information to the corresponding regions by matching the luminance and texture. Various correspondence techniques [29–33] have been proposed to improve the result of local color transfer by hand-crafted low-level features. Still, these correspondence methods are not robust to complex appearance variations of the same object because low-level features do not capture semantic information. With the development of deep learning, recent studies [34, 35] composed semantically close source-reference pairs based on features extracted from pretrained networks [34, 35] and exploited their semantic correlation for colorization. Although they may have good performance for photo colorization, their performance in sketch colorization is relatively weak. This is due to the abstract nature of sketches, which cannot offer enough visual semantic cues for dense color propagation, as widely used in photo colorization.

2.5 Image style-content disentanglement

We may also achieve reference-guided colorization by style-content disentanglement, in which images are dismantled into a *content* space and a *style*

3.2.2 Color style extractor

First, we use a color style extractor to extract a diversity of style variations in our training data into a unified color style space representation. The design of the color style space aims to collect as wide a range of different color styles as possible while excluding structural information from the style representation.

We use a 4-block downscaling residual network [38] as the style extractor. We input the style guidance image into the style extractor, and use global average pooling to output a style code with a dimension of 256. Our fully convolutional style extraction can extract the style code from a reference image of any resolution, but for convenience of training, we resize the image to 256×256 before input to the style extractor. We can combine the extracted color style features with any sketch to generate a new color cartoon image.

3.2.3 Colorization network with multi-scale outputs

The colorization network takes a cartoon sketch as input, colors it according to the color style code extracted from the reference image, and outputs a color cartoon with content consistent with the input and style consistent with the reference. Our colorization network is based on the U-Net structure, with a downscaling sketch encoder and an upscaling style-content fusion decoder. The encoder transforms a sketch image into deep feature maps with rich semantic information. The decoder fuses the color style code with the high-level feature maps of the input sketch to align the color style of the output to the reference. Unlike existing colorization networks that usually produce only one output image, we use a multi-scale output mechanism to help to train the discriminator to better distinguish realistic cartoon

images in terms of both global and local color characteristics. In particular, low-resolution output images help more to improve global color composition, while high-resolution images help more to improve local color shading.

In the encoder, we use 5 levels of downsampling blocks to obtain a hierarchical feature map with rich semantic information about the input sketch. We add instance normalization [39] to the encoder to ensure sketch style information erasure [7]. The feature map is then fed to the decoder. The decoder contains 5 upsampling blocks. We use concatenation operations to propagate information from the encoder to the decoder for better synthesis and reconstruction. The final output has the same resolution as the input sketch. In the decoder, we use AdaIN layers [7] to control the statistics of the feature map to achieve style editing and alignment to the reference image. After each upsampling block of the decoder, we produce an extra output image of a lower resolution, which is fed to the multi-scale discriminator and is also useful in constraining the loss function of our system; see Sections 3.2.4 and 3.3.2.

3.2.4 Multi-scale discriminator

We further employ a multi-scale discriminator to regularize the colorization network, for more realistic cartoon image generation. Unlike the commonly used discriminator network that judges generation quality based on a single input, we allow the discriminator to be compatible with different resolutions of generator output, vastly improving the receptive field of discrimination. We find that such a design leads to better colorization quality, especially in terms of global color composition. Figure 3 shows an example using single-scale and multi-scale discriminators. In

Fig. 3 Comparison with and without multi-scale outputs and losses.

Fig. 3(c), using a single scale only, the discriminator fails to capture the global color composition of the output image, in particular the colors of the legs and the eyes.

We construct our discriminator with 3 downscaling residual blocks. The discriminator is patch-based [1] and we compute the mean of the output as the discriminator output. The adversarial objective will be introduced in Section 3.3.2.

3.3 Training

3.3.1 Training dataset

To train our networks, we use a publicly available dataset [40] which contains 17,769 pairs of color cartoons and corresponding sketches. We use 14,224 pairs for training and 3545 for evaluation.

During training, for each color cartoon and the corresponding sketch, we take the sketch as the input image and feed it to the colorization network. Then we take the color cartoon as the reference image and feed it to the color style extractor. Ideally, the output image should take the content from the sketch and take the color style from the color cartoon, i.e., the output image should be the same image as the color cartoon.

3.3.2 Loss function

Our loss function contains two loss terms, a multi-scale reconstruction loss and a multi-scale adversarial loss.

The *reconstruction loss* ensures the functionality of style extraction and style propagation by determining the reconstruction ability of the colorization network. This is done by computing the difference between a ground-truth color cartoon and the network output by the color cartoon as style guidance and its sketch counterpart as input. By minimizing the reconstruction error, the network can better learn the style encoding in a more precise way and provide a color output similar to the ground truth. As mentioned before, our colorization network provides multi-scale versions of the colored output, so we employ our reconstruction loss at each level of the output so that the network can learn colorization and style propagation in a coarse-to-fine manner and balance the learning load for all upscaling convolutions. Specifically, for each output resolution, the reconstruction loss depends on the perceptual

loss [16] and the pixel-wise mean square error (MSE) via a weighted sum:

$$\mathcal{L}_{\text{rec}} = \frac{1}{n} \sum_i^n \lambda_i \|\varphi(\hat{y}_i) - \varphi(y_i)\|_2^2 + \frac{\omega_1}{n} \sum_i^n \lambda_i \|\hat{y}_i - y_i\|_2^2 \quad (1)$$

Here, $\hat{y}_i$ is the image predicted for a certain level generated by the colorization network. y_i is the ground-truth image of the same resolution, obtained by rescaling the original resolution ground-truth image using bilinear interpolation. λ_i is the weight for each level of output, set to [1, 2, 3, 10] respectively; output images with higher resolutions have higher weights. $\varphi(\cdot)$ is the output of the VGG16 network [41]. ω_1 is a weight set to 5 in all our experiments.

We further adopt a *multi-scale adversarial loss* [42] for our multi-scale discriminator. To further improve stability, we apply gradient penalty regularization [43] to the discriminator. Our multi-scale adversarial loss is defined as

$$\mathcal{L}_{\text{adv}} = \frac{1}{n} \sum_i^n \lambda_i E_{y_i} [\min(0, -D(y) - 1)] + \frac{1}{n} \sum_i^n \lambda_i E_{\hat{y}_i} [\min(0, D(\hat{y}_i) - 1)] + \omega_2 E_y [\|\nabla D(y)\|^2] \quad (2)$$

Here, D is the discriminator. $E[\cdot]$ is the expectation operator. ω_2 weights the gradient penalty and is set to 10 in all our experiments. The multi-scale adversarial loss increases the realism of the colorization results in terms of both global color composition and local color shading, as shown in Fig. 3. It also widens the diversity of different types of style references.

The overall loss function $\mathcal{L}$ of our framework is the sum of the reconstruction loss $\mathcal{L}_{\text{rec}}$ and the multi-scale adversarial loss $\mathcal{L}_{\text{adv}}$:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{adv}} \quad (3)$$

3.3.3 Training details

We use the Adam optimizer [44] to train our networks. All networks are jointly trained. The learning rate is initially set to 10^{-4} and gradually decreased to 2×10^{-6} . We employ a learning rate adjustment policy where the initial learning rate is multiplied by $1 - (\text{iter}/\text{max_iter})^{0.9}$. The optimization converges after about 150 epochs: see Fig. 4.

Fig. 4 Training loss for individual loss terms and total loss.

4 Results

In this section, we present an in-depth evaluation of our reference-guided sketch colorization framework. First, we present visual comparisons between our method and several competitors in different categories to qualitatively evaluate the performance of colorization and style alignment of our framework. We further perform a quantitative comparison to mainstream state-of-the-art colorization methods in terms of our sketch colorization task. Moreover, we present an ablation study to investigate the design and contribution of each component in our framework.

We categorize our competitors into three major categories: image style transfer, style-content disentanglement, and conditional sketch colorization. We choose several state-of-the-art works in each category as the benchmark. For image style transfer, we choose three recent CNN-based methods, the Gatys method [15], WCT [8], and AdaIN [7] as our competitors. For style-content disentanglement, we compare with CCD [9] and DMIT [10]. Both methods are trained with our prepared cartoon dataset. For conditional sketch colorization, we choose two state-of-the-art reference-based colorization methods [5, 27] and one state-of-the-art hint-based colorization method [25] as our competitors.

4.1 Qualitative evaluations

Figure 5 visually compares our method and state-of-the-art image style transfer, style-content disentanglement, and reference-based sketch colorization methods. Gatys' method [15] fails to colorize the sketch and only splashes random color patterns. AdaIN [7] and WCT [8] achieve better results, but structural lines are not well preserved with obvious distortions and artifacts. The style-content

disentanglement methods preserve the structural lines of the sketch better, but the color styles of the generated images are usually dissimilar to those of the reference cartoon. This is mainly because they encode the content component via an implicit representation, which may introduce bias to the shape and introduce extra errors when encoding the color style. Similarly, the reference-based colorization method [27] fails to propagate the exact color from the reference during colorization. In sharp contrast, our method preserves the content component by faithfully reproducing the structural lines in the sketch, so that both the style and the content are represented in our colorization framework with less bias.

Figure 6 visually compares our method and state-of-the-art hint-based sketch colorization methods. As shown in Fig. 6(c), the results of reference-based colorization [5] contain unexpected color mixing and obvious color discontinuity. Moreover, the results do not fully reflect the color characteristics of the reference image. The hint-based colorization method in Ref. [25] produces a style similar to the reference image given a certain number of user hints (at least 25 color points), as shown in Figs. 6(d) and 6(e). However, the quality of the output highly depends on the user's expertise in color composition and requires extensive color hints. The colorization results often deviate from the reference image and the overall coloring procedure may be inconvenient for amateur users. In comparison, our method faithfully propagates the reference color styles to the sketches with much less user effort.

Besides, we also investigated using reference-based colorization methods from other domains to our sketch colorization task. Icon colorization [45] and photo colorization [35] were tested. A visual comparison is presented in Fig. 7. We can observe that Ref. [45] learns to propagate the colors from the reference, but the output colors are too saturated for real-life sketch colorization purposes. Ref. [35] failed to obtain semantic correspondence in our sketch colorization task and thus outputs very dull colorization results. The domain gap is too large to directly adopt photo and icon colorization techniques for cartoon sketches.

4.2 Quantitative evaluation

We next present a quantitative evaluation based on the quality of reconstruction. We randomly sampled

Fig. 5 Visual comparison between our method and state-of-the-art image style transfer methods, style-content disentanglement methods, and conditional sketch colorization.

ground-truth cartoon and sketch pairs from the evaluation dataset and used the color cartoon as the style guidance to colorize the sketch. Ideally, the colored output image should be exactly the same as the color cartoon. To estimate the reconstruction quality, we measured the similarity between the reconstructed color image and the ground-truth color image using two commonly used similarity metrics, peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [46]: see Table 1. Our method

outperforms all competitors in both metrics in the perspective of reconstruction.

Furthermore, to estimate the realism of our colorized outputs, we also present a quantitative study by calculating the Fréchet inception distance (FID) [47] between our colorization results and ground-truth color cartoons. In this case, the style reference and the input sketch do not need to be the same. Again see Table 1: our method produces the closest colorization results to the ground truth, again

Fig. 6 Visual comparison between our method and state-of-the-art conditional sketch colorization methods.

demonstrating the superiority of our method in terms of output realism.

4.3 Ablation study

To validate the impact of each component in our network design, we performed an ablation study, determining the reconstruction metrics and FID metrics with different network designs. See Table 2;

note that the PSNR/SSIM metrics are evaluated with paired sketch/reference input and FID is evaluated with random references.

Firstly, without the multi-scale output design, both PSNR and SSIM values dropped significantly, which shows the effectiveness of our multi-scale output design. We also studied the use of the multi-scale discriminator by only feeding a single full-resolution

Table 1 Quantitative evaluation

Method	PSNR	SSIM	FID
Gatys	13.53	0.54	187.56
WCT	15.41	0.59	161.93
AdaIN	14.07	0.64	81.56
DMIT	12.33	0.69	46.95
CCD	13.61	0.69	104.85
Lee's	16.00	0.76	50.12
Ours	17.25	0.74	27.99

Table 2 Ablation study

Method	PSNR	SSIM	FID
Without multi-scale outputs	18.89	0.81	33.50
With multi-scale discriminator	19.99	0.84	36.80
With multi-scale reconstruction	22.01	0.84	32.35
AdaIn→feature reshape & concat	20.37	0.68	66.48
Full method with 128D style	21.88	0.84	30.38
Full method with 512D style	19.52	0.72	39.82
Our full method (256D style)	22.43	0.85	27.99

output to the discriminator. The results showed that the multi-scale discriminator design is essential to the output quality. The multi-scale discriminator design brings together the adversarial learning at different scales and can be seen as a general case of Ref. [48], which fuses global and local information together to improve generated results. With the multi-scale network design, the colorization can take care of both global color composition and the local texture synthesis. In addition, we find that the multi-scale reconstruction loss is also very important for generating visually pleasant results. We also tested replacing adaptive instance normalization with an

alternative style injection approach in which style codes of the same spatial size are reshaped as feature maps and concatenated. However, as shown in the second row, both measurements dropped significantly with this concatenation design. We also investigated the appropriate dimension of code used to encode the color style. Higher-dimensional encoding has a larger capacity but may encode some extra information that may not be directly related to style. Also, using higher dimensions may reduce the generalizability of the style encoding. On the opposite side, lower-dimensional encoding has better generalization but may lead to a shift in color style in the reconstructed image. In our experiments, we found 256 dimensions to be optimal for the color style code, best balancing reconstruction and generalization.

We also explored the effectiveness of the discriminator by visual assessment of an ablation, as shown in Fig. 8. Without the discriminator, the diversity of color in the output image can be restricted, and the color styles can be overfitted. For example, in the first row, without adversarial learning, the colorization of the girl's eyes loses symmetry, so appears unnatural. The same issue is also observed in the last row, where the network without adversarial learning incorrectly propagates white colors to the top-right of the canvas due to similar modes in the style reference. Adversarial learning minimizes the impact of these style-unrelated modes and ensures natural colorization in these cases. On the other hand, adversarial learning also effectively alleviates the problem of color-bleeding, e.g., on the girl's chest

Fig. 7 Comparison of our method to icon colorization and photo colorization methods.

Fig. 8 Comparison: with and without the discriminator.

in the second row and the girl's arms and hair in the third row.

5 Limitation

Although our method can be applied to most sketch images, our output results may still be affected by the color and texture composition of the style reference. If the style reference contains very little color information, the output quality may not be satisfying. Additionally, if the style reference and the input sketch are too different in structural composition, our method may find difficulties in propagating color styles from the colored regions of the reference style to the input sketch, as shown in Fig. 9.

Fig. 9 Our method may not work well when the style reference contains very few color information or when the style reference and the input sketch are too different in structure composition.

6 Conclusions

In this paper, we proposed a novel deep learning approach for reference-guided cartoon sketch colorization. Our system consists of a color style extractor that extracts a color style code from a color cartoon image, a colorization network that fuses a sketch with a color style code to generate a set of multi-scale color cartoon outputs, and a multi-scale discriminator that improves the realism of the generated color cartoon in terms of both global color composition and local color shading. Experiments show that our method significantly outperforms existing methods in content preservation of the input and style consistency with the reference images. As future work, we intend to explore the potential of colorizing an animation sequence by incorporating temporal information.

Acknowledgements

This work was supported in part by a CIHE Institutional Development Grant No. IDG200107, the National Natural Science Foundation of China under Grant No. 61973221, and the Natural Science Foundation of Guangdong Province of China under Grant Nos. 2018A030313381 and 2019A1515011165.

References

- [1] Isola, P.; Zhu, J. Y.; Zhou, T. H.; Efros, A. A. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5967–5976, 2017.
- [2] Zhu, J. Y.; Park, T.; Isola, P.; Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision*, 2242–2251, 2017.
- [3] Zhang, R.; Isola, P.; Efros, A. A. Colorful image colorization. In: *Computer Vision—ECCV 2016. Lecture Notes in Computer Science, Vol 9907*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 649–666, 2016.
- [4] Yonetsuji, T. Paints chainer. 2017. Available at <https://github.com/pfnet/PaintsChainer>.
- [5] Zhang, L.; Ji, Y.; Lin, X.; Liu, C. P. Style transfer for anime sketches with enhanced residual U-net and auxiliary classifier GAN. In: *Proceedings of the 4th IAPR Asian Conference on Pattern Recognition*, 506–511, 2017.
- [6] 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, 9055–9064, 2019.
- [7] Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, 1510–1519, 2017.
- [8] Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; Yang, M. Universal style transfer via feature transforms. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 385–395, 2017.
- [9] Gonzalez-Garcia, A.; van de Weijer, J.; Bengio, Y. Image-to-image translation for cross-domain disentanglement. In: Proceedings of the 33rd Conference on Neural Information Processing Systems, 1294–1305, 2018.
- [10] Yu, X.; Chen, Y.; Liu, S.; Li, T.; Li, G. Multi-mapping image-to-image translation via learning disentanglement. In: Proceedings of the 33rd Conference on Neural Information Processing Systems, 2990–2999, 2019.
- [11] Li, X. J.; Zhao, H. L.; Nie, G. Z.; Huang, H. Image recoloring using geodesic distance based color harmonization. *Computational Visual Media* Vol. 1, No. 2, 143–155, 2015.
- [12] Miao, Y. W.; Hu, F. X.; Zhang, X. D.; Chen, J. Z.; Pajarola, R. SymmSketch: Creating symmetric 3D free-form shapes from 2D sketches. *Computational Visual Media* Vol. 1, No. 1, 3–16, 2015.
- [13] Todo, H.; Yamaguchi, Y. Estimating reflectance and shape of objects from a single cartoon-shaded image. *Computational Visual Media* Vol. 3, No. 1, 21–31, 2017.
- [14] Hertzmann, A.; Jacobs, C. E.; Oliver, N.; Curless, B.; Salesin, D. H. Image analogies. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, 327–340, 2001.
- [15] Gatys, L. A.; Ecker, A. S.; Bethge, M. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2414–2423, 2016.
- [16] Johnson, J.; Alahi, A.; Li, F. F. Perceptual losses for real-time style transfer and super-resolution. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9906*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 694–711, 2016.
- [17] Sanakoyeu, A.; Kotovenko, D.; Lang, S.; Ommer, B. A style-aware content loss for real-time HD style transfer. In: Proceedings of the European Conference on Computer Vision, Vol. 11212, 715–731, 2018.
- [18] Zhang, Y. L.; Fang, C.; Wang, Y. L.; Wang, Z. W.; Lin, Z.; Fu, Y.; Yang, J. Multimodal style transfer via graph cuts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 5942–5950, 2019.
- [19] Song, C.; Wu, Z.; Zhou, Y.; Gong, M.; Huang, H. ETNet: Error transition network for arbitrary style transfer. In: Proceedings of the Advances in Neural Information Processing Systems, 668–677, 2019.
- [20] Wang, H.; Li, Y. J.; Wang, Y. H.; Hu, H. J.; Yang, M. H. Collaborative distillation for ultra-resolution universal style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1857–1866, 2020.
- [21] Li, X.; Liu, S.; Kautz, J.; Yang, M. Learning linear transformations for fast arbitrary style transfer. *arXiv preprint arXiv: 1808.04537*, 2018.
- [22] Gao, W.; Li, Y. J.; Yin, Y. H.; Yang, M. H. Fast video multi-style transfer. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 3211–3219, 2020.
- [23] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Communications of the ACM* Vol. 63, No. 11, 139–144, 2020.
- [24] Kim, T.; Cha, M.; Kim, H.; Lee, J. K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, 1857–1865, 2017.
- [25] Zhang, L.; Li, C. Z.; Wong, T. T.; Ji, Y.; Liu, C. P. Two-stage sketch colorization. *ACM Transactions on Graphics* Vol. 37, No. 6, Article No. 261, 2019.
- [26] Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; Shechtman, E. Toward multimodal image-to-image translation. In: Proceedings of the 31st Conference on Neural Information Processing Systems, 465–476, 2017.
- [27] 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, 5800–5809, 2020.
- [28] Welsh, T.; Ashikhmin, M.; Mueller, K. Transferring color to greyscale images. In: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, 277–280, 2002.
- [29] Bugeau, A.; Ta, V. T.; Papadakis, N. Variational exemplar-based image colorization. *IEEE Transactions on Image Processing* Vol. 23, No. 1, 298–307, 2014.

- [30] Liu, X. P.; Wan, L.; Qu, Y. G.; Wong, T. T.; Lin, S.; Leung, C. S.; Heng, P. A. Intrinsic colorization. In: Proceedings of the ACM SIGGRAPH Asia 2008 papers, Article No. 152, 2008.
- [31] Chia, A. Y. S.; Zhuo, S. J.; Gupta, R. K.; Tai, Y. W.; Cho, S. Y.; Tan, P.; Lin, S. Semantic colorization with Internet images. *ACM Transactions on Graphics* Vol. 30, No. 6, Article No. 156, 2011.
- [32] Gupta, R. K.; Chia, A. Y. S.; Rajan, D.; Ng, E. S.; Huang, Z. Y. Image colorization using similar images. In: Proceedings of the 20th ACM International Conference on Multimedia, 369–378, 2012.
- [33] Tai, Y. W.; Jia, J. Y.; Tang, C. K. Local color transfer via probabilistic segmentation by expectation-maximization. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 747–754, 2005.
- [34] He, M. M.; Chen, D. D.; Liao, J.; Sander, P. V.; Yuan, L. Deep exemplar-based colorization. *ACM Transactions on Graphics* Vol. 37, No. 4, Article No. 47, 2018.
- [35] Zhang, B.; He, M. M.; Liao, J.; Sander, P. V.; Yuan, L.; Bermak, A.; Chen, D. Deep exemplar-based video colorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8052–8061, 2019.
- [36] Huang, X.; Liu, M. Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11207*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 179–196, 2018.
- [37] Lee, H. Y.; Tseng, H. Y.; Huang, J. B.; Singh, M.; Yang, M. H. Diverse image-to-image translation via disentangled representations. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11205*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 36–52, 2018.
- [38] He, K. M.; Zhang, X. Y.; Ren, S. Q.; Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778, 2016.
- [39] Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.
- [40] Available at <https://www.kaggle.com/ktaebum/anime-sketch-colorization-pair>.
- [41] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [42] Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral normalization for generative adversarial networks. In: Proceedings of the 6th International Conference on Learning Representations, 2018.
- [43] Mescheder, L.; Geiger, A.; Nowozin, S. Which training methods for GANs do actually converge? In: Proceedings of the 35th International Conference on Machine Learning, 3478–3487, 2018.
- [44] Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations, 2015.
- [45] Sun, T. H.; Lai, C. H.; Wong, S. K.; Wang, Y. S. Adversarial colorization of icons based on contour and color conditions. In: Proceedings of the 27th ACM International Conference on Multimedia, 683–691, 2019.
- [46] Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* Vol. 13, No. 4, 600–612, 2004.
- [47] Dowson, D. C.; Landau, B. V. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis* Vol. 12, No. 3, 450–455, 1982.
- [48] Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. *ACM Transactions on Graphics* Vol. 36, No. 4, Article No. 107, 2017.

Xueting Liu received her B.Eng. degree in computer science and technology from Tsinghua University and her Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong in 2009 and 2014, respectively. She is currently an assistant professor in the School of Computing and Information Sciences, Caritas Institute of Higher Education. Her research interests include computational art, intelligent art, computer vision, and computer graphics.

Wenliang Wu received his B.Sc. degree from Guangdong Ocean University of Science and Technology in 2019. He is currently a graduate student in the College of Computer Science and Software Engineering, Shenzhen University. His research interests include computer vision and computer graphics.

Chengze Li received her B.Eng. degree from the University of Science and Technology of China in 2013, and her Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong in 2020. Chengze is currently an assistant professor in the School of Computing and Information

Sciences, Caritas Institute of Higher Education, with research interests in 2D non-photorealistic media analysis and processing, computational photography, and computer graphics.

Yifan Li received his B.Sc. degree from Jiangxi University of Science and Technology in 2018 and is now a graduate student in the College of Computer Science and Software Engineering, Shenzhen University. His research interests include computer graphics, computer vision, machine learning, and deep learning.

Huisi Wu received his B.E. and M.E. degrees in computer science both from Xi'an Jiaotong University in 2004 and 2007, respectively. He obtained his Ph.D. degree in computer science from the Chinese University of Hong Kong in 2011. He is currently an associate professor in the College of Computer Science and

Software Engineering, Shenzhen University. His research interests include computer graphics, image processing, and medical imaging.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made.

The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Other papers from this open access journal are available free of charge from <http://www.springer.com/journal/41095>. To submit a manuscript, please go to <https://www.editorialmanager.com/cvmj>.