Received: 12 June 2023 | Revised: 23 June 2023 | Accepted: 4 July 2023

DOI: 10.1002/cav.2198

SPECIAL ISSUE PAPER

WILEY

Hand-drawn anime line drawing colorization of faces with texture details

Kenta Akita | **Yuki Morimoto** | **Reiji Tsuruno**

Kyushu University, Fukuoka, Japan

Correspondence
Kenta Akita, Kyushu University, Fukuoka, Japan.
Email: akita.kenta.633@s.kyushu-u.ac.jp

Funding information
Japan Science and Technology Agency, Grant/Award Number: JPMJSP2136

Abstract

Automatic or semi-automatic colorization can reduce the burden of illustrators in color illustration production, which is a research area with significant market demand. Texture details in eyes and hair influence the impression of character illustrations. Generally, these details are not expressed in line drawings. Many existing automatic or semi-automatic colorization methods do not target hand-drawn line drawings and it is difficult to paint texture details on such drawings. In this paper, we propose the semi-automatic colorization of character line drawings around faces with texture details. Our method uses a reference image as a color hint and transfers the textures of the reference image to a line drawing. To achieve this, our method uses semantic segmentation masks to match parts of the line drawing with the same parts of the reference image. We create two types of segmentation datasets to train a segmentation network that creates segmentation masks. We transfer texture details to a hand-drawn line drawing by mapping each part of the reference image to the corresponding part of the line drawing using segmentation masks. We show that our method is more effective for hand-drawn line drawings than existing methods using qualitative and quantitative evaluations.

KEYWORDS

colorization, deep learning, line drawings

1 | INTRODUCTION

Color character illustrations are used in various content, such as games and art books, and are in high demand in the content industry. In the coloring process for color character illustrations, illustrators colorize each part of line drawings, such as hair and skin, using flat color, followed by painting details, such as pupils, shading, and highlights. We collectively refer to the pupil and iris as pupils in this paper because they are considered together in color character illustration production. The colorization process is a time-consuming and labor-intensive task. Thus, automatic or semi-automatic colorization is important for effective color illustration production and for reducing labor (Figure 1).

Hand-drawn anime line drawings used in color illustration production (Figure 2 left) are drawn with contour lines and few lines in the region inside the pupils and hair.[1] Contour lines in this paper are the contour lines of each region in illustrations, such as hair, pupils, and skin. In color illustration production, textural details, such as pupil details, shading, and highlights, inside pupils and hair are painted in hand-drawn line drawings during colorization.

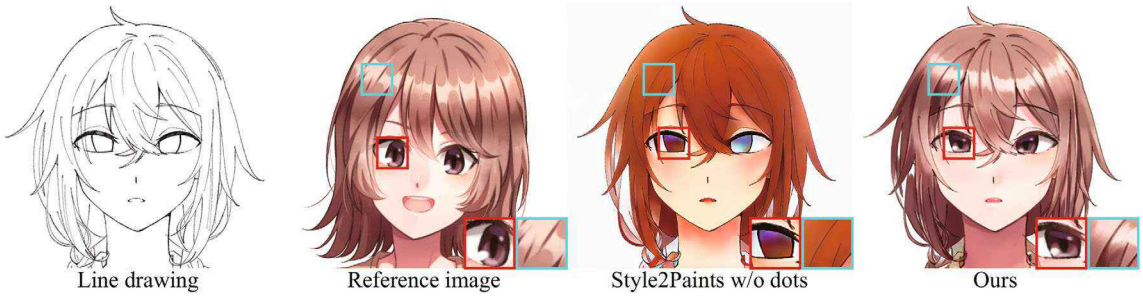

FIGURE 1 Comparison between the colorization images produced by Style2Paints[6] and our method. © kou kankitu (reference image).

FIGURE 2 Line drawing used in color illustration production (left) and a colorization image (middle), line drawing extracted using the sketch extraction method[7] (right). The extracted line drawing has edges of textural details.

These textural details are essential for influencing the impression of color illustrations. Many semi-automatic colorization methods[2-5] have been proposed recently; however, these methods cannot express textural details on hand-drawn line drawings. There is a reason that these methods train colorization networks using line drawings extracted from color images with edges such as surrounded by color rectangles in Figure 2 right. Because of such edges, the extracted line drawings have large domain gaps from the hand-drawn line drawings. Automatic or semi-automatic colorization, which can paint textural details on hand-drawn line drawings during colorization, would be useful in color illustration production.

In this paper, we propose reference-based semi-automatic colorization for hand-drawn anime face line drawings to represent the textural details of the reference image, for example, pupils, hair, and skin (Figure 1 right). We train a colorization network using line drawings with small domain gaps to transfer textural details from a reference image to a line drawing. Generally, textural details represented during colorization in color illustrations are different in each region: highlights and shading for hair, shading and blush for skin, and pupils and highlights for eyes. The correspondence between each part of the reference image and line drawing is very important for correctly transferring the textural details in each part. We create two types of semantic segmentation datasets to train a semantic segmentation network and use segmentation generated by the network to determine the correspondence. The colorization model trained using line drawings with small domain gaps can transfer textural details from a reference image to a hand-drawn line drawing using the semantic segmentation mask.

We use CoCosNet v2[8] as the colorization network. CoCosNet v2[8] creates a mapping image that maps color and textural details from a reference image to a segmentation mask and then synthesizes the image using the mapping image. We call the mapping image the hint image. To correctly transfer the textural details from the reference image to the line drawing, we train the colorization network using transformed reference images using image augmentation. During testing, the colorization network sometimes fails to determine the correspondence between the line drawing and reference image. In this case, the colorization network transfers color and textural details from the reference image to the wrong regions of the line drawing. We propose two training steps to suppress this problem and generate a reasonable colorization result.

We show that our colorization model achieves higher-quality images than state-of-the-art methods. We evaluate both the colorization results and the transfer of textural details. To evaluate the transfer of textural details, we use metrics to evaluate a line drawing and reference image with different shapes.

To summarize, the three main contributions of our study are as follows:

- We propose a semi-automatic colorization model that transfers textural details from a reference image to a hand-drawn line drawing using segmentation masks and small domain gap line drawings.
- We propose a two-step training method for the colorization model that is robust to correspondence failures in each region.
- We show that our colorization model outperforms state-of-the-art methods using multiple evaluation metrics.

2 | RELATED WORK

Our method transfers textural details from a reference image to a line drawing of a character illustration and colorizes the line drawing. These are similar to style transfer and image synthesis, respectively. Our method also creates segments that are used to colorize the line drawing. Therefore, we refer to related work on semi-automatic colorization in Section 2.1, image synthesis and style transfer in Section 2.2, and the segmentation mask dataset in Section 2.3.

2.1 | Learning-based colorization of line drawings

Most semi-automatic colorization methods use color hints to enable users to specify a color. These color hints are mainly scribbles,[2,6,7,9] text,[3] and reference images.[4,10,11] The semi-automatic colorization methods described above allow colorization that reflects user input. Although the methods can represent gradations of color in a hand-drawn line drawing, it is difficult to express textural details, such as color edge, highlights, shading, and shadows, in regions such as the pupils, skin, and hair.

Akita et al.[12] proposed semi-automatic colorization for hand-drawn anime line drawings to express details only on the pupils of character illustrations. Ashtari et al.'s method[13] transfers the style from a line drawing of a reference image to a color image. A colorization network can be trained on line drawings with a small domain gap using this method. Seg2pix[14] colorizes line drawings without hints, but the method must create a dataset for each character illustration and train a network with the dataset.

We propose semi-automatic colorization to enable the painting of textural details around the face, particularly shading and highlights, such as pupils, hair, and skin, using reference images.

2.2 | Image manipulation

Neural style transfer[15-17] transfers reference image details to the input image. The method is effective for color image to color image transfer, but does not work for line drawings to color image transfer. Image synthesis[18-22] using generative adversarial networks (GANs)[23] generates high-quality photorealistic synthesis images by providing segmentation masks, pose images, and sketches.

In addition to methods that use only label images, methods that use reference images to specify the style of the synthesized image have been proposed.[8,24,25] These methods can express texture details; however, in the case applied to illustrations, the dataset's preparation is difficult because many segmentation masks are required and the details represented in these methods are not sufficient. We create the two types of segmentation dataset and introduce the two-step training method, and solve these problems.

Recently, several image synthesis methods have been developed[26-28] based on diffusion models,[29,30] in which sketches or line drawings form the input. When hand-drawn line drawings are input to ControlNet,[26] however, there are often undesirable changes, such as lines being stuck together and a closed mouth becoming an open mouth. Although our method targets only the face, the lines of the input line drawing do not change.

2.3 | Face parsing

Semantic segmentation[31-33] segments objects in the image semantically and labels each object. The labels are per object and regions in an object cannot be divided. In the case of image synthesis per face part,[34,35] semantic part segmentation that target faces, often called face parsing, is required to segment each part of the object.

Datasets for semantic part segmentation and face parsing are the PASCAL-Part dataset[36] and CelebMask-HQ dataset.[34] However, a dataset for illustration targeted by our method does not exist.

In this paper, we create a face parsing dataset for illustrations and line drawings, and build an illustration segmentation model.

3 | OVERVIEW

Figure 3 shows an overview of our method in the test phase. In our method, the user inputs a reference image and line drawing. The segmentation network creates segmentation masks. Then, we use two segmentation models: one for the color image and one for the line drawing. In the segmentation network, the line drawing is transformed into a distance field image.

In the colorization network, a mapping network generates correspondences between each part of the reference image and the line drawing. This process uses the reference image, segmentation mask of the reference image, and segmentation mask of the line drawing. The mapping network then generates a hint image based on these correspondences. Finally, the translation network colorizes the line drawings using the hint image.

4 | DATASET PREPARATION

We create datasets to train the segmentation network and colorization network. The two types of segmentation dataset we use in training the segmentation network are the coarse dataset and fine dataset. The datasets we use to train the colorization network are the color character illustration dataset (I_{color}), segmentation mask dataset created from I_{color} using the segmentation model (S_{color}), line drawing dataset extracted from I_{color} (I_{line}), and segmentation mask dataset created from I_{line} using the segmentation model (S_{line}).

To create these datasets, we use Danbooru2021,[37] which is a large-scale anime illustration image database that includes character illustrations and the tags of the character illustrations. We select color character illustrations with Danbooru tags of the "simple background" and "1girl" or "1boy". We detect these images using a face detector and crop them. Then, to include hair, we crop the images from the detected face size to a size 1.5 times larger in the upward direction and 1.25 times larger in the left and right directions. The cropped images are larger than 512×512 pixels. After this process, we manually remove images, such as detection error images. Finally, we use approximately 37,000 images as the colorization dataset (I_{color}). Additionally, we creat a training dataset for the face parsing model from the dataset.

4.1 | Face parsing for anime illustration and line drawings

Segmentation masks are used to match to each region of a line drawing to a reference image. To train the segmentation network, creating the segmentation mask dataset for illustrations is necessary; however, it is very labor-intensive if it is

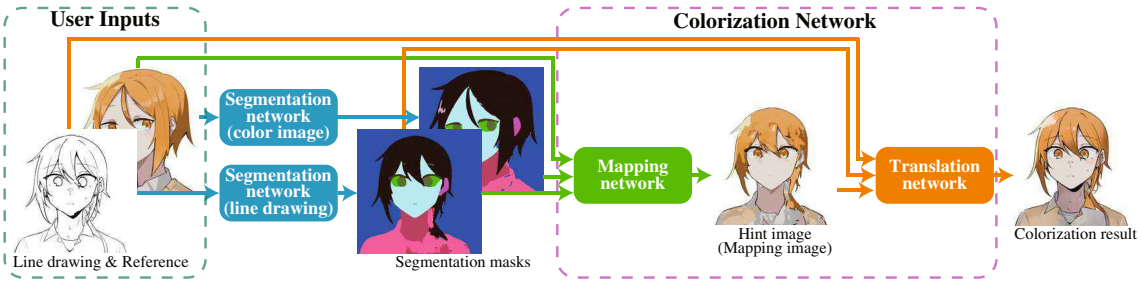

FIGURE 3 Overview of the proposed colorization system in the test phase.

created completely manually. Thus, we create a fine dataset and a coarse dataset and use these to reduce the need for annotation.

The fine dataset contains 220 images and segmentation masks, all of which are manually annotated. To reduce the labor-intensity of annotation, the coarse dataset is created by using the few-shot face parsing method.[38] The coarse dataset includes about 37,000 images and segmentation masks. Our method labels are 12 classes: background, clothes and accessories, upper body and hands, mouth, eyebrows, eyelashes, pupils, white of eyes, ears, face, hair, and animal ears (e.g., rabbit ears and cat ears). The images in the fine dataset are created by illustrators.

We create the coarse dataset using the few-shot face parsing method. This method uses pre-trained StyleGAN2,[39] which we re-train[1] on I_{color}. Few-shot Segmenter[38] generates segmentation masks from latent variables into which real images are embedded. Therefore, we project latent variables from I_{color} in our method. We manually annotate 10 images generated from the projected latent variables. We train the network using these segmentation masks and latent variables. Then, we select 5,000 generated images with a small loss from the projected latent variables and train the segmentation model on these generated images and segmentation masks. Finally, we create a coarse dataset from I_{color} using the segmentation model.

4.2 | Line extraction

We input hand-drawn line drawings into the colorization network; however, it is difficult to prepare many hand-drawn line drawings that correspond to the color images during training. Therefore, we use line drawings extracted from color images.

Line drawings extracted from color images using Xdog,[40] sketch extraction,[7] and sketchKeras,[41] which are often used in existing semi-automatic colorization approaches, contain the edges of textural details, but hand-drawn line drawings in color illustration production do not contain these edges (Figure 2 right). The extract line drawings are different from hand-drawn line drawings, which leads to large domain gaps. Thus, we extract line drawings from color images during training using the learning-based TOM.[21] TOM can extract line drawings that are close to hand-drawn line drawings, which means that there are small domain gaps.

5 | SEGMENTATION NETWORK

The inputs of the colorization network are a reference image, line drawing, and their segmentation masks. Our segmentation network architecture is the same as that of the auto-shot segmenter.[38] To generate the segmentation masks, we train the segmentation network models with the reference image and the line drawing transformed into a distance field image as inputs. We use I_{color} to train the segmentation model for color images and I_{line} to train the segmentation model for line drawings. we split the training of the segmentation network into two stages, each using a different dataset (Figure 4). We use the coarse dataset for the first training stage and the fine dataset for the second. The segmentation network trains the coarse semantic segmentation process in the first stage and the fine semantic segmentation process in the second stage. Using these two datasets to train the segmentation networks ensures that, although some manual annotation is required, relatively high segmentation accuracy can be achieved with fewer manual annotations than when training is performed using only one of the two datasets.

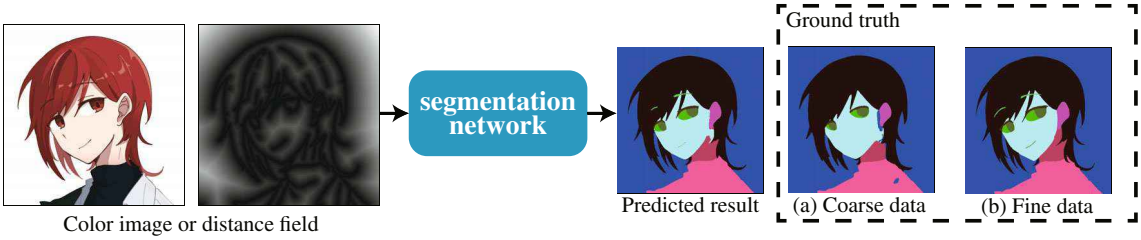

FIGURE 4 Training of the segmentation network. We trained the network on the dataset of (a), and then the dataset of (b).

[1]We use the model introduced at EXTENDED STYLEGAN2 DANBOORU2019, AYDAO https://www.gwern.net/Faces.

6 | COLORIZATION NETWORK

Our colorization network is the same as the CoCosNet v2 network,[8] which enables high-resolution image synthesis. The network inputs are a segmentation mask, reference image, and line drawing. The CoCosNet v2 network consists of a mapping network that corresponds to the segmentation mask and reference image, and creates a hint image (mapping image) based on the correspondence and a translation network that colorizes a line drawing using the hint image (Figure 5 top). The segmentation mask, hint image, and line drawing are concatenated and input into the translation network.

We split the training of the colorization network into two steps. In the first step, we train the translation and mapping network. In the second step, we only train the translation network using a hint image, similar to the test phase. We improve the colorization in the case of correspondence f ailures through two-step training.

Additionally, to create a segmentation mask, we use the segmentation network trained with only the coarse dataset in the training phase because the performance of the colorization model is improved, and we use the segmentation network trained with the two segmentation dataset types in the test phase. We use S_{color} instead of S_{line} in the first step using image augmentation so that a reference image can correspond easily to a line drawing.

6.1 | Image augmentation for training

The mapping network learns the semantic correspondence between each region of a segmentation mask of a line drawing and each region of a reference image. As input, the mapping network requires a reference image and segmentation masks of a line drawing and reference image. Ideal training datasets that consist of the same people or objects in different poses

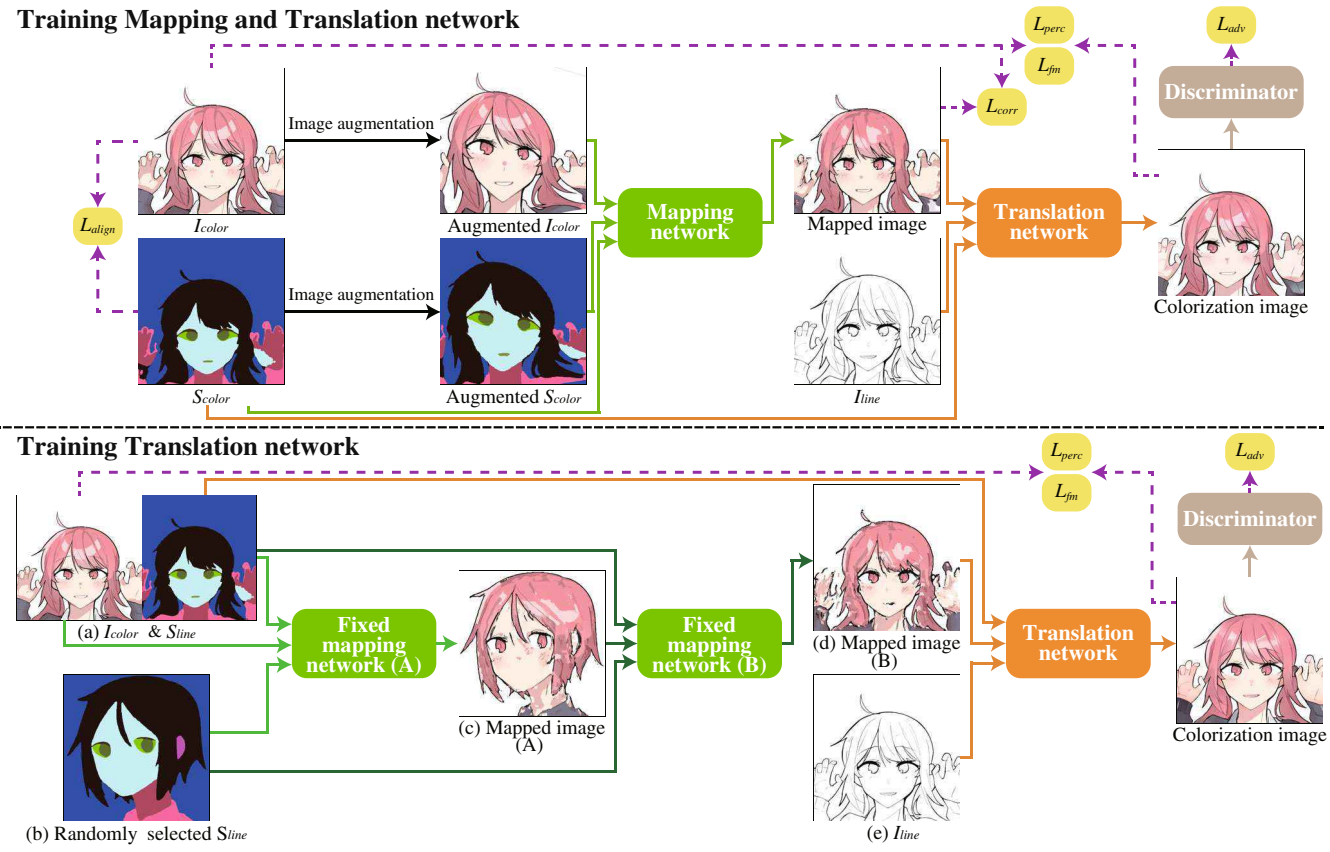

FIGURE 5 Training of the colorization network. The colorization network is trained in two steps. In the first step, a segmentation mask and a reference image with the TPS transformation, and a segmentation mask without the transformation are input into the mapping network. The fixed mapping network is the mapping network with fixed weights.

and directions do not exist. Thus, in the case of the photographed images, input images are used for training that are similar to the texture, shape, and alignment of each region. Similar images have high cosine similarity. Cosine similarity is calculated using the extracted features from images in a dataset using pre-trained VGG19.[42]

During character illustration, we train the colorization network on a reference image and a segmentation mask similar to the reference image. However, colorization network training with similar images only slightly transfers textural details from a reference image to a line drawing. When we train the colorization network using corresponding images to reflect color of a reference image, color representation improves, but textural details cannot be represented adequately. Therefore, we introduce the thin plate spline (TPS) transformation to solve these problems. When we create a hint image using the mapping network, input I_{color} and S_{color} are augmented by the TPS transformation. Then, we input the hint image, S_{color}, and I_{line} into the translation network and train the network (Figure 5 top). Augmentation solves the problems of the color reproduction and representation of textual details, and the translation network transfers the textural details of the reference image to the line drawing.

6.2 | Training for correspondence failures

As described previously, the mapping network creates a hint image using the correspondence between each region of a segmentation mask and reference image. The image is a hint image in the translation network. However, during testing, semantic face parts of input images are naturally different (Figure 3), which may cause the correspondence to fail. To suppress the influence of correspondence failure on colorization, this training method creates a hint image that includes such correspondence failures (Figure 5d) similar to the test phase. During this training, we do not apply the TPS transformation (Figure 5a,b) and do not train the mapping network.

In the following, we describe how to create the hint image (c) with reference to the bottom of Figure 5. First, we input I_{color} (a), the corresponding S_{line}, and a segmentation mask randomly selected from S_{line} (b) into the first fixed mapping network (A) and map I_{color} to the non-corresponding S_{line} (c) to create an image. Next, the fixed mapping network (B) creates an image (d) that maps the mapped image (c) to S_{line} (a) that corresponds to I_{color}. Then, we do not update the weights of the mapping network and update the weights of the translation network. We train the translation network on the hint images (d), the corresponding I_{line} (e), and S_{line} (a), and the translation network processes the corresponding failures in the test phase.

7 | EXPERIMENTS

7.1 | Implementation details

Segmentation network. When we train the segmentation model, we use the coarse dataset for training only, and split the fine dataset in the ratio of 9:1 for training and validation. We train the segmentation model for approximately 29,000 iterations with a batch size of 64 on the first dataset and for approximately 5200 iterations with a batch size of 16 on the second dataset. We use the Adam solver with $\beta_1 = 0.9$ and $\beta_1 = 0.99$. The learning rate were $lr = 0.0008$ at first training and $lr = 0.0002$ at second training. We set a decay factor of 0.1 within 650 iterations for training on the fine dataset.

Colorization network. We train the colorization network in two steps and chang the weights of the loss functions in the first and second steps. We use the loss functions of CoCosNet v2[2] as the loss functions of the colorization network, given by

$$L_{total} = w_1 L_{align} + w_2 L_{corr} + w_3 L_{perc} + w_4 L_{fm} + w_5 L_{adv},$$

where L_{align} is the domain alignment loss, L_{corr} is the correspondence loss, L_{perc} is the perceptual loss, L_{fm} is the feature matching loss, L_{adv} is the adversarial loss, and w is the weight for each loss function. In the first step, we set the weight for each loss function as follows: $w_1 = 10$, $w_2 = 500$, $w_3 = 0.001$, $w_4 = 10$, and $w_1 = 50$. In the second step, we set the weight for each loss function as follows: $w_1 = 0$, $w_2 = 0$, $w_3 = 0.001$, $w_4 = 10$, and $w_1 = 50$. We train the colorization model for

[2]The loss functions used are shown in the published source code. https://github.com/microsoft/CoCosNet-v2/blob/main/models/pix2pix_model.py.

approximately 37,000 iterations, with a batch size of 16 for both stages of training. We use the Adam solver with $\beta_1 = 0.5$ and $\beta_1 = 0.999$. The learning rate were 4e−4 for both the generator and discriminator.

7.2 | Qualitative results

Figure 6 shows the results of our method, existing semi-automatic colorization methods, and existing image synthesis methods. All output images are colorized based on the reference color image. Tag2Pix[3] uses tags as color hints. These tags specify regions with specified colors, such as red hair. The Tag2Pix method enables colorization with few inputs, but the specified color is inflexible. The scribble-based and reference-base methods[4,6,7] colorize the line drawing similar to the reference image's color. However, whereas the scribble-based method can represent hair highlights, to some extent, the above methods are difficult for them to represent textural details, such as pupils and hair shading. The method developed by Lee et al. (Figure 6d) only reflects the color of hair and skin in the reference image. Although the method of Akita et al.[12] (Figure 6f) can represent the pupil details of the reference image, it sometimes fails to predict the pupil positions. SPADE[24] (Figure 6g) does not use a line drawing as input, but instead takes a segmentation mask. This method produces good texture details, but the color differs from that of the reference image. The method of Liu et al.[21] (Figure 6h) uses a line drawing rather than a segmentation mask. The method accurately represents the color in the reference image, but does not represent the pupils of the reference image and produces strong artifacts. Our method colorizes (Figure 6i) the line drawing close to the color of the reference image and rarely has artifacts. The proposed approach represents the style of aspects such as the pupils, hair highlights, and shading in greater detail than existing methods.

Figure 7 shows the results of our method and existing style transfer methods. The method of Gatys et al.[15] transfers only the global style of the reference image, whereas Deep Image Analogy[16] transfers both the global and local styles of the reference image. Gatys et al.'s method is not effective for line drawings, and Deep Image Analogy cannot represent hair highlights or pupils. By contrast, our method transfers the textural details of each part of the reference image to the corresponding part of the line drawing.

7.3 | Quantitative results

To evaluate the colorization results, we compare the performance of our method with that of SPADE, Lee et al.'s method, Akita et al.'s method, and Liu et al.'s method. We use the Fréchet inception distance (FID),[43,44] peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and learned perceptual image patch similarity (LPIPS)[45] as evaluation metrics. Additionally, we use the modified cyclic evaluation metric.[13]

FIGURE 6 Comparison of existing colorization methods and image synthesis methods. Color hints for each method are (a) color scribbles,[7] (b) color dots,[6] (c) natural language,[3] (d) the method of Lee et al.,[4] (e) Style2Paints without dots,[6] (f) the method of Akita et al.,[12] (g) SPADE,[24] (h) the method of Liu et al.,[21] and (i) color reference images. All output images are colorized based on the reference color image. © yuyu ringo (reference image).

FIGURE 7 Comparison of existing style transfer methods.[15,16] © kou kankitu (reference image).

For each evaluation, except for FID, we use the results of line drawings colorized by each method as inputs and color illustrations of the same line drawings colorized by illustrators as the ground truths. Because the FID evaluations require many images, we use 59 hand-drawn line drawings and 67 reference images to create 3953 (= 59× 63) color images. We then obtain the multivariate normal distribution of the colorization results. We obtain the distribution of the ground truth from I_{color}.

To conduct the PSNR, SSIM, LPIPS, and cyclic evaluations, we collect 42 hand-drawn line drawings and the corresponding color illustrations from online sources. The color illustrations are used as the ground truths. We colorize the line drawings using each method, and then use the ground truths as the reference images for each metric, except for the cyclic evaluation. We do not apply image augmentation to the reference images so as not to lose the texture details of the reference images. Segmentation masks are generated from the line drawings and the reference images. Table 1 shows the evaluation results of our method and the existing methods. The evaluation results show that our method is the best in all categories.

The inputs of Ashtari et al.'s method is a hand-drawn line drawing as a reference image and a color image, and a line drawing is extracted from the color image. The cyclic evaluation process evaluates the performance of each method in terms of the PSNR when we use reference images that do not correspond to the line drawings. This is because this metric transfers the texture details to line drawings with different shapes. The original cyclic evaluation procedure evaluates line drawings extracted from color illustrations. However, we wish to evaluate the performance of each method transferring the texture details of a reference image to a line drawing. Therefore, we change the input to line drawings and the output to color images and introduce a suitable constraint. We set the constraint as follows:

$$C_{ref}^{label} \subset L_{in}^{label},$$

where C_{ref}^{label} is labels of reference images and L_{in}^{label} is labels of input line drawings.

This constraint limits the loss of texture details during the generation of a segmentation mask for a line drawing that does not correspond to the reference image; for example, if there is a cloth region on one side but not on the other. This enables an accurate evaluation of the transfer performance of a model. Additionally, when there is too much difference between the shape of the line drawing and the shape of the reference image, the model is more likely to fail to transfer the textural details from the reference image to the line drawing. Thus, we colorize the line drawing using the top-5 most-similar images according to the cosine similarity (Section 6.1) as reference images.

TABLE 1 Comparison of the evaluation metrics for the proposed method and existing methods.

Method	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Cyclic-PSNR ↑	Cyclic-SSIM ↑	Cyclic-LPIPS ↓
SPADE[24]	29.50	13.21	0.431	0.385	12.32	0.416	0.397
Lee et al.[4]	98.25	13.73	0.599	0.352	12.96	0.581	0.345
Akita et al.[12]	53.91	15.39	0.623	0.351	14.61	0.603	0.321
Liu et al.[21]	27.45	14.14	0.550	0.377	13.03	0.517	0.421
Our method	**16.89**	**17.15**	**0.742**	**0.231**	**15.52**	**0.675**	**0.274**

7.4 | User study

To evaluate our method, we conducted an online questionnaire with sixteen participants. We compared our method to Liu et al.'s method and Akita et al.'s method. We arranged the colorization images for each method, and participants ranked these images from 1 to 3.

We created 300 colorization images (10 line drawings × 30 reference images) for each method and displayed 10 sets selected from these images. The line drawings and reference images we used for each method were the same in each set. We randomly selected and randomly arranged the colorization results. The evaluation items were as follows:

- **Quality of color illustrations**
  How good is the quality of the colorization image as a color illustration?
- **Reflected textural details of reference images**
  How well are the color and details, for example, pupils and hair highlights, in the reference image reflected in the colorization image?

When we question the item of reflected textural details of reference images, we simultaneously displayed the line drawing and reference image in addition to the colorization images.

Table 2 shows the questionnaire results. In the user study on the quality of color illustrations, our method was ranked first for most of the questions, which indicates that it created more complete color illustrations than the existing methods.

7.5 | Ablation study

Segmentation network. To validate the improvement in the segmentation accuracy of color illustrations and the distance fields into which line drawings are transformed when using the coarse and fine datasets, we compare the segmentation model trained using the two datasets (coarse and fine) with segmentation models trained on either the coarse or fine dataset. We use the same evaluation metrics used for Tritrong et al.'s method[38]: intersection over union (IOU) per part and weighted IOU (WIOU). WIOU is defined as follows:

$$WIOU = \sum_c \left(\frac{S_c}{S_{all}} IOU_c \right),$$

where c is the class, S_{all} denotes all pixels, and S_c denotes the ground-truth pixels that belong to the class.

To create segmentation masks of the ground truths, we annotate 40 hand-drawn line drawings and the images of these line drawings colorized by humans. The evaluation results (Table 3) demonstrate the segmentation accuracy achieved by performing network training on both datasets, with the highest scores in many items.

The segmentation accuracy of the network for distance field training on the coarse dataset scores highest for animal ears and upper body, and the segmentation accuracy of the network for distance field training using the fine dataset scores highest for brows. This is because these regions have various appearances or are either simply present or absent.

TABLE 2 User study on the quality of color illustrations (left) and the reflection of textural details (right).

Rank	Liu et al.[21] (%)	Akita et al.[12] (%)	Our method (%)
Top 1	8.75	10.0	**81.3**
Top 2	44.4	47.5	8.13
Top 3	46.9	42.5	10.6
Top 1	8.8	10.6	**80.6**
Top 2	64.4	28.1	7.5
Top 3	26.9	61.3	11.9

TABLE 3 WIOU and IOU for the face parsing of color images and line drawings.

Dataset	Clothes	Ears	Animal ears	Upper body & hands	Face	Mouth	Brows	Eyes	Lashes	White of eyes	Hair	Background	WIOU
Fine (color)	50.0	19.6	4.8	60.9	88.1	57.8	38.1	80.9	59.5	59.2	82.7	90.3	80.7
Coarse (color)	40.7	22.8	0.3	53.0	87.8	31.6	0.0	83.8	44.0	40.1	83.0	91.0	79.2
Full (color)	**59.3**	**49.0**	**7.5**	**67.9**	**92.4**	**62.5**	**41.3**	**86.5**	**65.5**	**69.2**	**88.1**	**95.9**	**86.9**
Fine (distance)	46.2	18.3	4.5	36.5	83.0	49.0	**38.4**	82.7	56.4	50.8	77.8	79.2	74.8
Coarse (distance)	46.1	26.8	**10.3**	**53.5**	88.4	32.1	0.0	83.8	48.5	40.5	78.0	84.7	76.2
Full (distance)	**56.8**	**36.2**	8.0	48.4	**91.2**	**56.0**	27.3	**87.7**	**62.2**	**60.0**	**84.0**	**85.8**	**81.8**

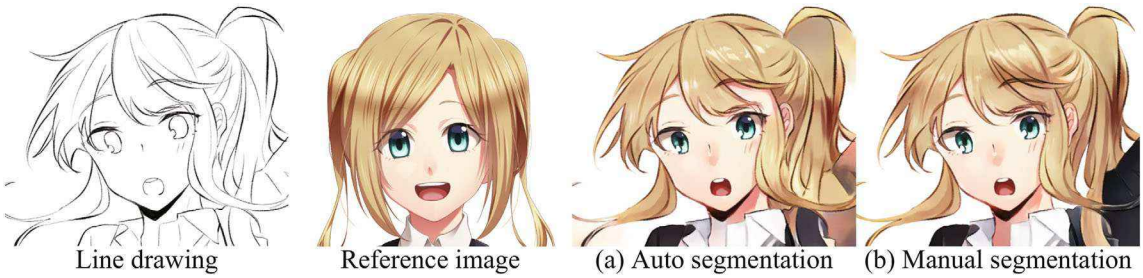

FIGURE 8 Difference between the colorization results caused by the segmentation accuracy: (a) colorization result with automatic segmentation and (b) colorization result with manually modified segmentation. © kou kankitu (reference image).

The segmentation accuracy of the face region is higher than that of other regions. The segmentation masks in the fine dataset are ideal because they are annotated color illustrations created by illustrators. However, illustrations generated by StyleGAN2[39] are used to create the coarse dataset, and the illustrations are of low quality except for the faces. As a result, even after using both datasets to improve the accuracy, the problem remains to some extent because the fine dataset is relatively small. Therefore, the accuracy of the segmentation network is low except for face regions.

Segmentation accuracy influences the colorization results. Our method sometimes fails to produce a reasonable segmentation. In such cases, the colors are reflected in incorrect regions and color bleeding occurs. For example, the cloth regions are colorized by the color of the hair in the reference image in Figure 8a. They are improved by the manual modification of the segmentation mask, as shown in Figure 8b. Manual modification can be simplified using an interactive user interface such as Delaunay Painting.[46]

Colorization network. To improve the transfer of the textural details of the reference images to the line drawings, we introduced image augmentation using the TPS transformation, proposed a training method that suppressed the effects of correspondence failures, and trained the networks using line drawings with small domain gaps. To confirm these effects, we evaluated the methods with and without image augmentation, with and without the training method for correspondence failures, training using similar images, and training using a mixture of extracted line drawings with large domain gaps using XDoG[40] and sketch extraction[7] using FID, PSNR, SSIM, LPIPS, and cyclic evaluation.

Table 4 shows the evaluation results. The metrics, except for FID, showed little difference for all methods. FID was superior for colorization networks with augmentation without two-step training and the full method. When the positions of texture details are misalignment, scores of SSIM, PSNR, and LPIPS are low. In this case, these scores of images with blurred texture details are higher. FID can evaluate the quality of colorization results. Therefore, we select the model with the highest FID.

Figure 9 shows a comparison of the full method with the method without image augmentation and the method using similar images. When we trained the colorization network without image augmentation, the network represented the textural details of the reference image, but the colorization results were unnatural. The colorization network trained using similar images hardly transfers the color and textural details of the reference image.

Figure 10 compares the full method with the method without two-step training. The colorization results were similar, but the full method suppresses the influence of correspondence failure on the colorization result. Figure 11 shows the

TABLE 4 Evaluation of the methods with and without image augmentation, with and without the training method for correspondence failures, training using similar images, and training using line drawings with large domain gaps.

Method	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Cyclic-PSNR ↑	Cyclic-SSIM ↑	Cyclic-LPIPS ↓
Without augmentation	48.50	17.45	0.728	0.216	15.48	0.615	0.333
Training using similar images	28.04	**17.70**	**0.770**	**0.198**	14.87	0.677	0.282
Augmentation without two-step training	19.69	17.44	0.761	0.205	15.62	0.685	0.263
Training using line drawings with a large domain gap	24.23	17.50	0.769	0.218	**15.83**	**0.710**	**0.254**
Augmentation with two-step training	**16.89**	17.15	0.742	0.231	15.52	0.675	0.274

FIGURE 9 Comparison of training without TPS transformation and training using similarity images. © kou kankitu (reference image).

FIGURE 10 The mouth in the line drawing is open, but the mouth in the reference image is closed, which indicates poor correspondence. © mizuharu (reference image).

FIGURE 11 Comparison of training with line drawings with large and small domain gaps. © mizuharu (reference image).

colorization results for each colorization network trained using line drawings with large and small domain gaps. The colorization network trained using line drawings with small domain gaps transfers the textural details from the reference image to the line drawing better than the colorization network trained using line drawings with large domain gaps.

8 | CONCLUSION

In this paper, we propose a semi-automatic colorization method that can transfer textural details from a reference image to a hand-drawn line drawing used in color illustration production. The segmentation network segments anime character illustrations with high accuracy using two types of segmentation datasets. We use TPS to reflect the textural details of the reference image and improve the colorization quality using two-step training to suppress correspondence failures. The qualitative and quantitative experimental results show that our colorization model outperforms state-of-the-art methods. We create a segmentation mask dataset that segmented anime character face illustrations into parts, and propose a segmentation model for the illustrations.

ACKNOWLEDGMENTS

This work was supported by JST SPRING, Grant Number JPMJSP2136 and was partly achieved through the use of SQUID at the Cybermedia Center, Osaka University.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Kenta Akita https://orcid.org/0009-0000-7300-123X

REFERENCES

1. Hart C. *The Master Guide to Drawing Anime: Amazing Girls: How to Draw Essential Character Types from Simple Templates*. 2017.
2. Zhang L, Li C, Wong TT, Ji Y, Liu C. Two-stage sketch colorization. ACM Trans Graph. 2018;37:6.
3. Kim H, Jhoo HY, Park E, Yoo S. Tag2Pix: Line art colorization using text tag With SECat and changing loss. In: International Conference on Computer Vision (ICCV). New York, NY: IEEE; 2019.
4. 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 (CVPR). New York, NY: IEEE; 2020. p. 5801–10.
5. Cao R, Mo H, Gao C. Line art colorization based on explicit region segmentation. Comput Graph Forum. 2021;40(7):1–10.
6. lllyasviel. style2paints V4.5. https://github.com/lllyasviel/style2paints; 2018.
7. Yonetsuji T. Petalica paint. https://petalica-paint.pixiv.dev/index_en.html; 2017.
8. Zhou X, Zhang B, Zhang T, et al. CoCosNet v2: Full-resolution correspondence learning for image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2021. p. 11465–75.
9. Ci Y, Ma X, Wang Z, Li H, Luo Z. User-guided deep anime line art colorization with conditional adversarial networks. ACM Int Conf Multimedia (MM). 2018;MM '18:1536–44.
10. Furusawa C, Hiroshiba K, Ogaki K, Odagiri Y. Comicolorization: Semi-automatic Manga Colorization. New York, NY: SIGGRAPH Asia Technical Briefs; 2017.
11. Cui J, Zhong H, Liu H, Fu Y. Exemplar-based sketch colorization with cross-domain dense semantic correspondence. Mathematics. 2022;10(12):1–19.
12. Akita K, Morimoto Y, Tsuruno R. Colorization of line drawings with empty pupils. Comput Graph Forum. 2020;39(7):601–10.
13. Ashtari A, Seo CW, Kang C, Cha S, Noh J. Reference based sketch extraction via attention mechanism. ACM Trans Graph. 2022;41(6):1–16.
14. Seo CW, Seo Y. Seg2pix: Few shot training line art colorization with segmented image data. Appl Sci. 2021;11(4):1–16.
15. Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2016. p. 2414–23.
16. Liao J, Yao Y, Yuan L, Hua G, Kang SB. Visual attribute transfer through deep image analogy. ACM Trans Graph. 2017;36(4):1–15.
17. Li B, Belongie S, Sn L, Davis A. Neural image recolorization for creative domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New York, NY: IEEE; 2022. p. 2225–9.
18. Isola P, Zhou T, Zhou T, Efros AA. Image-to-Image translation with conditional adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2017. p. 5967–76.
19. Chen W, Hays J. SketchyGAN: Towards diverse and realistic sketch to image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2018. p. 9416–25.

20. Chen SY, Su W, Gao L, Xia S, Fu H. DeepFaceDrawing: Deep generation of face images from sketches. ACM Trans Graph. 2020;39(4):1–16.
21. Liu B, Zhu Y, Song K, Elgammal A. Self-supervised sketch-to-image synthesis. Proc AAAI Conf Artificial Intell. 2021;35(3):2073–81.
22. Wang M, Yang GY, Li R, et al. Example-guided style-consistent image synthesis from semantic labeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2019. p. 1495–504.
23. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Neural Informat Process Syst (NeurIps). 2014;27:2672–80.
24. Park T, Liu MY, Wang TC, Zhu JY. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2019. p. 2337–46.
25. Zhan F, Yu Y, Wu R, et al. Bi-level feature alignment for versatile image translation and manipulation. In: Proceedings of the European Conference on Computer Vision (ECCV). Berlin, Heidelberg: Springer; 2022. p. 224–41.
26. Zhang L, Agrawala M. Adding Conditional Control to Text-to-Image Diffusion Models. 2023.
27. Voynov A, Abernan K, Cohen-Or D. Sketch-Guided Text-to-Image Diffusion Models. 2022.
28. Mou C, Wang X, Xie L, et al. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. 2023.
29. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Neural Informat Process Syst (NeurIps). 2020;33:6840–51.
30. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2022. p. 10684–95.
31. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2015. p. 3431–40.
32. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). Berlin, Heidelberg: Springer; 2018. p. 801–18.
33. Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Analy Mach Intell (TPAMI). 2017;39(12):2481–95.
34. Lee CH, Liu Z, Wu L, Luo P. MaskGAN: Towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2020. p. 5549–58.
35. Zhu P, Abdal R, Qin Y, Wonka P. SEAN: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2020. p. 5104–13.
36. Chen X, Mottaghi R, Liu X, Fidler S, Urtasun R, Yuille AL. Detect What You Can: Detecting and representing objects using holistic models and body parts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2014. p. 1979–86.
37. Anonymous, Community D,Branwen G, Gokaslan A. Danbooru2021: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset. https://www.gwern.net/Danbooru2021 2022.
38. Tritrong N, Rewatbowornwong P, Suwajanakorn S. Repurposing GANs for one-shot semantic part segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2021. p. 4475–85.
39. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2020. p. 8110–9.
40. WinnemöLler H, Kyprianidis JE, Olsen SC. Xdog: An extended difference-of-Gaussians compendium including advanced image stylization. Comput Graph. 2012;36(6):740–53.
41. lllyasviel sketchkeras. https://github.com/lllyasviel/sketchKeras 2018.
42. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (ICLR); 2015.
43. 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. Neural Informat Process Syst (NeurIps). 2017;30:6629–40.
44. Parmar G, Zhang R, Zhu JY. On aliased resizing and surprising subtleties in GAN evaluation. CVPR. New York, NY: IEEE; 2022.
45. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2018. p. 586–95.
46. Parakkat AD, Memari P, Cani MP. Delaunay painting: Perceptual image colouring from raster contours with gaps. Comput Graph Forum. 2022;41(6):166–81.

AUTHOR BIOGRAPHIES

Kenta Akita received B.S. and M.S. degrees in design from Kyushu University, Japan, in 2019 and 2021, respectively. His research interests are in computer graphics and deep learning.

Yuki Morimoto received a Ph.D. degree in design from Kyushu University, Japan, in 2008. She is an assistant professor at Kyushu University. Her research interests are in computer graphics.

Reiji Tsuruno received a Ph.D. degree in engineering from Osaka Prefecture University, Japan, in 1987. He is a professor at Kyushu University. His research interests are in computer graphics and interaction.

How to cite this article: Akita K, Morimoto Y, Tsuruno R. Hand-drawn anime line drawing colorization of faces with texture details. Comput Anim Virtual Worlds. 2023;e2198. https://doi.org/10.1002/cav.2198