AnimeDiffusion: Anime Diffusion Colorization

Yu Cao, Xiangqiao Meng, P. Y. Mok, *Member, IEEE*, Tong-Yee Lee, *Senior Member, IEEE*, Xueting Liu, *Senior Member, IEEE*, and Ping Li, *Member, IEEE*

Abstract—Being essential in animation creation, colorizing anime line drawings is usually a tedious and time-consuming manual task. Reference-based line drawing colorization provides an intuitive way to automatically colorize target line drawings using reference images. The prevailing approaches are based on generative adversarial networks (GANs), yet these methods still cannot generate high-quality results comparable to manually-colored ones. In this paper, a new AnimeDiffusion approach is proposed via hybrid diffusions for the automatic colorization of anime face line drawings. This is the first attempt to utilize the diffusion model for reference-based colorization, which demands a high level of control over the image synthesis process. To do so, a hybrid end-to-end training strategy is designed, including phase 1 for training diffusion model with classifier-free guidance and phase 2 for efficiently updating color tone with a target reference colored image. The model learns denoising and structure-capturing ability in phase 1, and in phase 2, the model learns more accurate color information. Utilizing our hybrid training strategy, the network convergence speed is accelerated, and the colorization performance is improved. Our AnimeDiffusion generates colorization results with semantic correspondence and color consistency. In addition, the model has a certain generalization performance for line drawings of different line styles. To train and evaluate colorization methods, an anime face line drawing colorization benchmark dataset, containing 31,696 training data and 579 testing data, is introduced and shared. Extensive experiments and user studies have demonstrated that our proposed AnimeDiffusion outperforms state-of-the-art GAN-based methods and another diffusion-based model, both quantitatively and qualitatively.

Index Terms—Line drawing colorization, diffusion models, reference-based colorization, conditional GAN.

1 Introduction

Line drawing colorization is an essential process in the animation industry. However, manual colorization is time-consuming, especially for the content with line drawings of complex structures. It is therefore necessary and valuable to develop methods and systems for colorizing line drawings automatically. The task of line drawing colorization is indeed challenging because line drawings, different from grayscale images, contain only structure content composed of a series of lines without any luminance or texture information. This topic has caught great attention among researchers in the field of computer graphics and numerous approaches [1], [2], [3], [4] have been proposed for manga and cartoon line drawing colorization.

Among various kinds of approaches, the reference-based method [5] provides a convenient and effective way for colorization, in which users only need one reference color image, then the input line drawings can be automatically colorized *without any manual intervention*. It has a similar model structure to the exemplar-based image editing task [6], [7], [8] but requires essentially different model performance. Reference-based line drawing colorization was usually formulated as a conditional image generation task, and the prevailing approaches are mostly based on generative adversarial networks (GANs) [9], [10], [11]. The key endeavor of these conditioned-GAN methods lies in the feature aggregation of two extracted deep features from line drawings and reference color images. Nevertheless, GAN-based models suffer from the drawbacks of vanishing gradient, mode collapse, and unstable convergence.

Recently, diffusion probabilistic models [12] ("diffusion models" hereafter) have made encouraging progress in image synthesis tasks, especially in text-to-image generation [13], outperforming GANs in terms of output image quality and the diversity of the generated content [14], [15]. Many studies focus on generating diverse-quality images with a high level of control, leading to AI-generated content (AIGC) technology, by taking advantage of the outstanding generative capability of diffusion models. In terms of *conditional image synthesis*, it is still an active research topic in diffusion models, and existing strategies [8], [16] are mostly based on Stable Diffusion Model (SD) [13] pre-trained on multi-modal data. SD-based methods embed information from pixel space to latent space, allowing a certain level of control by multi-modal fusion using refined text prompts, i.e., *text-guided*.

In this paper, we propose a diffusion model tailored for reference-based anime face line drawing colorization, called **AnimeDiffusion**. We demonstrate a new way for conditional image synthesis using a diffusion model, through our

- *Yu Cao is with the School of Fashion and Textiles, The Hong Kong Polytechnic University, Hong Kong. E-mail: yu-daniel.cao@connect.polyu.hk.*
- *Xiangqiao Meng and Ping Li are with the Department of Computing and the School of Design, The Hong Kong Polytechnic University, Hong Kong. E-mail: xiangqiao.meng@connect.polyu.hk, p.li@polyu.edu.hk.*
- *P. Y. Mok is with the School of Fashion and Textiles, The Hong Kong Polytechnic University, Hong Kong, and also with the Laboratory for Artificial Intelligence in Design, Hong Kong. E-mail: tracy.mok@polyu.edu.hk.*
- *Tong-Yee Lee is with the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan 70101, Taiwan. E-mail: tonylee@ncku.edu.tw.*
- *Xueting Liu is with the School of Computing and Information Sciences, Caritas Institute of Higher Education, Hong Kong. E-mail: tliu@cihe.edu.hk.*

Manuscript received 02 May 2023; revised 26 December 2023.
This work was supported in part by the Innovation and Technology Commission of Hong Kong under Grant ITP/028/21TP, in part by the National Science and Technology Council under Grant 110-2221-E-006-135-MY3, Taiwan, and in part by The Hong Kong Polytechnic University under Grants P0030419, P0048387, P0042740, P0035358, P0043906, and P0044520.
Yu Cao and Xiangqiao Meng contributed equally to this work.
(Corresponding Authors: P. Y. Mok and Ping Li.)

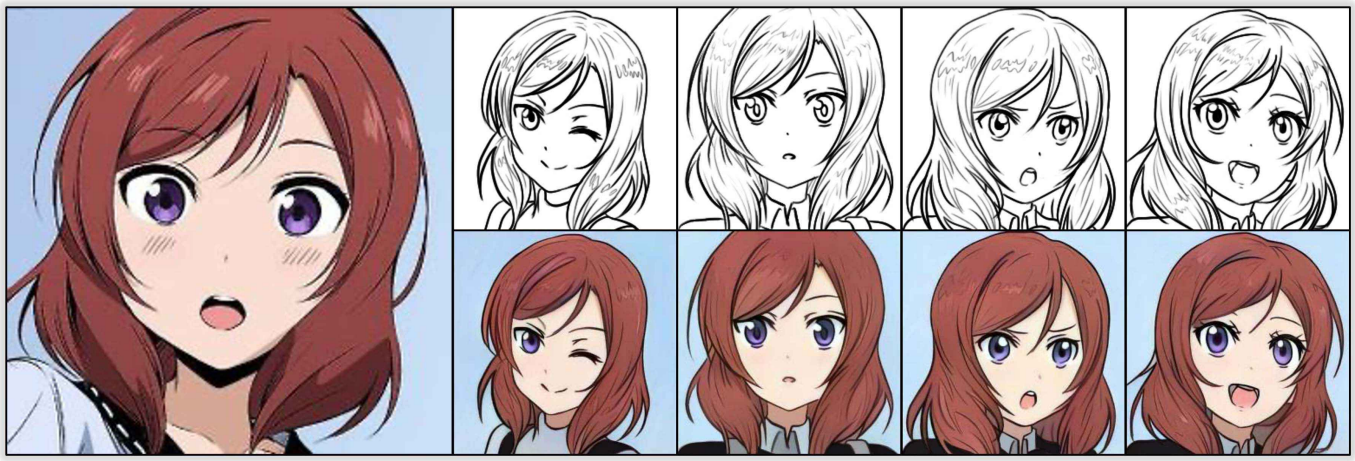


Fig. 1. AnimeDiffusion performs reference-based line drawing colorization: Given a reference color image (on the left) and four line drawings (on the top) of the anime character Nishikino Maki, it automatically generates colored results with accurate color and semantic correspondence. From left to right, the character's mouth is open, and the shape changes greatly, AnimeDiffusion can accurately colorize the mouth. In addition, the colored hair is richly layered and full of detailed textures. The eye color is accurate, and can accurately retain the pupil's white luster. We can also keep the background color consistent according to the reference image.

specific task of reference-based colorization, which demands a *high level of control over the generated results*, in terms of color consistency, spatial precision, and semantic correspondence [17]. Without relying on a specially designed fusion module or pre-trained text-guided diffusion model, our AnimeDiffusion can generate colored results with both high quality and diversity. We propose a **hybrid** training strategy through training task separation, and the training process consists of two phases, each with specific focus and loss. Phase 1 aims to train the model with good denoising ability, capturing global structure information, and the aim of phase 2 is to train the model for color information recovery ability. Different from other work for image-to-image translation [18], [19] or text-to-image translation [13], [16] that pre-trained text-guided diffusion models are fine-tuned, we train a diffusion model from scratch, on our own dataset, in two phases to perform colorization more efficiently, speed up the sampling process, and improve the quality of generation results. The hybrid training strategy also accelerates the convergence speed of the network. As shown in Fig. 1, our trained AnimeDiffusion can generate rich textures to the hair with detailed shading, instead of simply flat color, according to the input line drawings hand painted by an artist. Experimental results have demonstrated that AnimeDiffusion can generate better results than the state-of-the-art GAN-based line drawing colorization models and SD-based methods, both qualitatively and quantitatively. Despite the lack of available high-resolution anime face dataset, we build a new benchmark dataset from [20] and share for academic research at https://xq-meng.github.io/projects/AnimeDiffusion.

Our main contributions are summarized as follows:

- We develop the first diffusion model tailored for reference-based anime face line drawing colorization, i.e., AnimeDiffusion. It notably outperforms other GAN-based and SD-based models and achieves state-of-the-art anime face line drawing colorization results.
- We design a hybrid end-to-end training strategy for AnimeDiffusion, accelerating the convergence of network training as well as improving the colorization performance. This strategy effectively alleviates the problem that diffusion models are expensive to train and enables the model to obtain high-quality generation results within a limited number of inference steps.
- We establish a new benchmark dataset of paired line drawings and colored images of anime faces, which contains 31,696 training data and 579 testing data. It fills the gap of high-resolution (256×256) anime face dataset for training and evaluation.

2 Related Work

2.1 Line Drawing Colorization

Traditional line drawing colorization approaches [1], [2] are commonly optimized-based which allow users to use brushes to inject desired color into specific regions. Since line drawing contains only structure information with sparse line sets, existing automated colorization methods [21], [22], [23], [24] for grayscale images cannot be applied directly for line drawing colorization. Hence, numerous colorization methods tailored for line drawings have been developed.

Varga *et al.* [3] introduced the first deep learning-based method that colorizes cartoon images automatically with random colors. In order to have more effective control of the colored results, different user-hint-based methods were later proposed. The color hints are usually concatenated with line drawing and encoded as the input for neural networks in many deep learning-based methods. Ci *et al.* [25] proposed a conditional GAN model to colorize anime line drawings using color scribbles, which can generate colored results with accurate shading. Zhang *et al.* [26] developed a two-stage colorization method, which divided the entire colorization task into two simpler subtasks with clearer goals. Kim *et al.* [27] specially designed SECat module to generate illustrations with details using text tags as their hints. Zou *et*

al. [28] for the first time presented a language-based system for interactive colorization of scene sketches. However, such user-hints methods would become more labor-intensive as the number of line drawings increases, many interactions are needed to refine the colored results hence making these methods not user-friendly for amateur users without aesthetic training, especially when preparing appropriate color hints. Instead, many reference-based colorization methods have therefore developed.

Reference-based methods are very suitable for colorizing line drawing sets or videos of anime characters and the same characters require color consistency across frames. Lee *et al.* [9] proposed an attention-based Spatial Correspondence Feature Transfer (SCFT) module. Li *et al.* [10] eliminated the gradient conflict among attention branches by using the Stop-Gradient Attention (SGA) module. Cao *et al.* [11] designed an attention-aware model for generating high-quality colored anime line drawing images. Chen *et al.* [4] proposed an active learning based framework to match local regions between line art and reference colored image, followed by a mixed-integer quadratic programming (MIQP) that used spatial contexts to further refine matching results. Li *et al.* [29] proposed a learning method for sketch-driven cartoon video inbetweening. Shi *et al.* [30] proposed a new line art video colorization method using the 3D convolutional module to refine the temporal consistency of the colored results. Inspired by the previous human face editing work [31], [32], in this paper we focus on anime face line drawing colorization. Our AnimeDiffusion will be compared with the state-of-the-art GAN-based methods [9], [10], [11] in a later experiment section.

2.2 Diffusion Models

Diffusion models such as denoising diffusion probabilistic models (DDPM) [33] have achieved great success in image generation tasks. Diffusion-based image generation models show better performance than GAN-based models in terms of training stability and output image quality [14], [15]. Denoising diffusion implicit models (DDIM) [34] accelerated the sampling procedure and enabled a determined generation process with a given Gaussian noise. In addition to generating high-quality images based on random noise, diffusion models also perform well in conditional image-to-image translation tasks. An image-to-image diffusion model (Palette) [35] offered a versatile and general framework for image manipulation. Stochastic Differential Editing (SDEdit) [18] was a guided image editing and synthesis method, which synthesized realistic images by iterative denoising through a stochastic differential equation (SDE). Combined with the contrastive language-image pre-training (CLIP) model [36], diffusion models also support multi-modal generation tasks. DALL-E 2 [37] and Imagen [38] were proposed successively to perform text-guided natural image synthesis tasks. DiffusionCLIP [19] was designed to perform text-driven zero-shot image manipulation. Latent diffusion models [13], also called Stable Diffusion, can be trained using limited computational resources because of powerful autoencoders pre-trained in the latent space. Compared with GAN-based models, image generation based on diffusion models can easily add a variety of guidance, such as texts, strokes, and reference images. The existing diffusion models are designed for natural image generation with random noise or text prompts and some of them can perform natural image colorization based on prior color knowledge. However, these methods cannot be directly applied in the current task. Until very recently, Zhang *et al.* [16] proposed ControlNet that can generate a diversity of colored images including cartoons, according to the input sketch and text prompt. Nevertheless, since pre-trained Stable Diffusion [13] was used in ControlNet, a known drawback is that it cannot faithfully keep all line drawing details and sometimes produces exaggerated and distorted results according to input line drawings. In contrast, since we build and train the model from scratch, our approach offers a solution for greater control in line structure details and color texture preservation.

3 AnimeDiffusion

3.1 Overview

Given a reference image, we aim to colorize any input line drawings with clear geometry structure and accurate semantic colors. The core problem to solve is how to inject color from the corresponding position of the reference image into the line drawings. The training process of AnimeDiffusion consists of two phases. During phase 1, we concatenate the line drawing with a distorted reference image and the original reference image with noise together as input to our model. The phase 1 training is namely a **conditional denoising training** process, through which our model learns the basic noise removal capability and captures the global structure and correspondence between line drawing and color information. Nevertheless, the network may not accurately recover the color information, if purely based on denoising training, or it will take longer time to train because color space is huge. Therefore, for this colorization task, we train the model in phase 2 using **image reconstruction** loss, to quickly converge to the accurate color feature. Our AnimeDiffusion has two advantages: first, it eliminates the need for designing feature extractors, like other diffusion model-based methods for conditional generative tasks; and second, the loss functions of our diffusion model are simple and closely related to the training task. In the following, we will introduce model architecture, line extraction, and training strategy in detail.

3.2 Model Architecture

As illustrated in Fig. 2(a), assuming I_{gt} is an original colored image, from which a line drawing I_l is extracted by XDoG extractor [39]. Large spatial structure discrepancies between line drawings and reference images can be expected for any colorization task. To allow AnimeDiffusion to learn accurate semantic correspondence during the training process, we convert I_{gt} to a geometry-distorted version I_r by TPS transformation [40]. The forward diffusion converts I_{gt} to $I_{\text{noise}}^{(t)}$ with a random t steps in the range of T. $I_{\text{noise}}^{(t)}$, I_l and I_r are then concatenated together to comprise I_{concate} with 7 channels. We build a U-Net to predict the noise with 3 channels that are added to the I_{gt}. We propose a new conditional noise prediction proxy task in phase 1

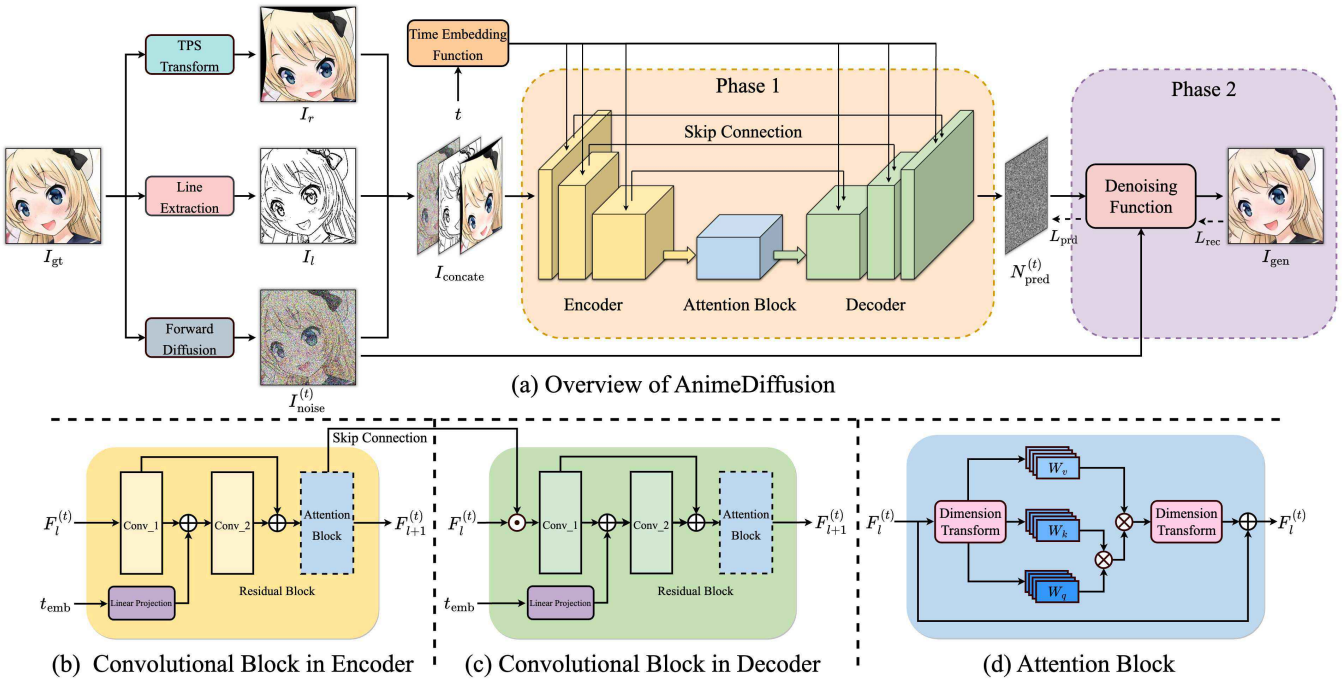


Fig. 2. The training flowchart of AnimeDiffusion. The top region shows the detailed structure of the noise prediction model and our designed hybrid training strategy. The bottom area shows a detailed encoder, decoder, and attention block structure. Our proposed hybrid training strategy separates the image denoising task and image reconstruction task, which improves the training efficiency of the network and makes the model have better coloring performance. Without introducing an additional discriminator network, the output images of the AnimeDiffusion have excellent coloring quality and the quality is very close to that of manually colored images.

training by introducing I_l and I_r as additional inputs of the condition. The information of t is embedded using the time embedding function and is transmitted to all convolutional blocks of both the encoder and decoder in the U-Net.

As is shown in Fig. 2(b) and Fig. 2(c), the convolutional blocks in the encoder and decoder have the same structure containing a residual block followed by an attention block. The dotted box is used to represent the attention block in the figure. Attention blocks are selectively included, in which they are excluded in the shallow layers of the encoder. The encoder and decoder equipped with multi-head self-attention enable AnimeDiffusion to more efficiently capture both global and local features in different convolutional layers. There is a linear projection module to map the embedded time information t_{emb} to feature maps of the same size after the first convolution operation in each layer. We use the common pixel-wise addition operation to encode the time information into the convolutional block. The attention block is not only used as a sub-block in the encoder and decoder of U-Net but also as an independent block in the bottleneck layer of U-Net. The detailed structure of the attention block is illustrated in Fig. 2(d). The use of an attention block allows the model to learn long-range features and multiscale features which are essential for the colorization task. Then, a denoising function is used to transfer the predicted noise $N_{\text{pred}}^{(t)}$ from input $I_{\text{noise}}^{(t)}$ so as to generate a colored image I_{gen}.

3.3 Line Extraction

Because of the lack of a large dataset for high-quality anime line drawings, we extract XDoG [39] line style, namely the intermediate representation of edge detection operator [39], is used as input for AnimeDiffusion training and inference. To do so, we incorporate a line extraction module to the AnimeDiffusion model (see Fig. 2). During the training stage, line drawings extracted from colored images are used as input to AnimeDiffusion, and during the inference stage, hand-drawn sketches of other line drawing input, such as artwork created by artists, are transformed to XDoG style as input. For given colored image I_{gt}, the line extractor to obtain I_l is described as [39],

$$S_{\sigma,k,p}(I_{\text{gt}}) = (1+p)\cdot G_{\sigma}(I_{\text{gt}}) - p\cdot G_{k\sigma}(I_{\text{gt}}), \tag{1}$$

where G_{σ} and $G_{k\sigma}$ are Gaussian convolution operations, σ is the variance of Gaussian convolution kernel, k is the scaling ratio of the variance between two convolutions, and p is used to control the edge emphasis lines [39].

We need a line extractor in which line extraction results are as close as possible to the effect of a painter's hand-drawn line. The variance σ of the Gaussian convolution has a significant effect on the line thickness of the line drawing, and we choose σ to be 0.3 to get a reasonable line width. We hired a professional artist to draw line art for some of the images in our dataset. For the colored image I_{gt}, its line drawing extraction result is $S_{k,p}(I_{\text{gt}})$, and the hand-drawn image by the artist is represented as $H(I_{\text{gt}})$. The objective of parameters for the line extractor is

$$\arg\min_{k,p} \sum_{i} \|S_{k,p}(I_{\text{gt}}) - H(I_{\text{gt}})\|_2. \tag{2}$$

3.4 Training Strategy

We design a hybrid training strategy to train AnimeDiffusion, which consists of a classifier-free guidance phase 1 training and an image reconstruction guidance phase 2 training. The hybrid also means we combine DDPM [33] and DDIM [34], i.e., DDPM for phase 1 training and DDIM for phase 2 training. This strategy separates the denoising task from the image reconstruction task, enabling the network to learn a specific task at each stage, which is beneficial for network training and weight updating.

3.4.1 Phase 1 Training

During the classifier-free guidance phase 1 training, AnimeDiffusion mainly learns denoising ability. As shown in Fig. 2, the original image I_{gt} goes through a forward diffusion process which is a Markov chain since it adds Gaussian noise to I_{gt} and obtains noisy image $I_{\text{noise}}^{(t)}$ for time step t iteratively. Each step of the forward process is a Gaussian transition,

$$q(I_{\text{noise}}^{(t)}|I_{\text{noise}}^{(t-1)}) = \mathcal{N}(I_{\text{noise}}^{(t)}; \sqrt{1-\beta_t}I_{\text{noise}}^{(t-1)}, \beta_t\mathbf{I}), \quad (3)$$

where β_t is variance schedule at time step t. The forward process of the diffusion model represents the addition of noise from step 0 to t. For the cumulative T steps of noise addition $I_{\text{noise}}^{(1:T)}$, the marginal distribution is

$$q(I_{\text{noise}}^{(1:T)}|I_{\text{gt}}) = \Pi_{t=1}^{T} q(I_{\text{noise}}^{(t)}|I_{\text{noise}}^{(t-1)}). \quad (4)$$

Under the condition of Eq. (3), the marginal distribution of each forward step is a standard Gaussian distribution

$$q(I_{\text{noise}}^{(t)}|I_{\text{gt}}) = \mathcal{N}(I_{\text{noise}}^{(t)}; \sqrt{\bar{\alpha}_t}I_{\text{gt}}, (1-\bar{\alpha}_t)\mathbf{I}), \quad (5)$$

where $\bar{\alpha}_t = \Pi_{i=1}^{t}(1-\beta_i)$. After t time steps, the result latent variable $I_{\text{noise}}^{(t)}$ can be simplified as

$$I_{\text{noise}}^{(t)} = \sqrt{\bar{\alpha}_t}I_{\text{gt}} + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (6)$$

The training objective of the model is to predict the noise $\epsilon_\theta(I_l, I_r, I_{\text{noise}}^{(t)}, t)$ with given noisy data point $I_{\text{noise}}^{(t)}$, time step t and condition I_l and I_r, and optimizing the objective

$$L_{\text{prd}} = \mathbb{E}_{I_{\text{gt}}\sim q(I_{\text{gt}}),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),I_l,I_r,t}\|\epsilon - \epsilon_\theta(I_l, I_r, I_{\text{noise}}^{(t)}, t)\|_2. \quad (7)$$

For simplicity, $\epsilon_\theta(I_l, I_r, I_{\text{noise}}^{(t)}, t)$ is abbreviated as ϵ_θ hereafter. We adopt the L2 norm in phase 1 training because the L2 norm could capture the output distribution more faithfully [35]. We used a model of U-Net with attention blocks to predict the noise ϵ added in Eq. (6). The U-Net needs to accept line drawing I_l, reference image I_r, and noisy image $I_{\text{noise}}^{(t)}$ as inputs. Considering the high spatial consistency between line drawing and color images, we concatenate the above three in the channel dimension. The experimental results in Section 4 demonstrate that, without a complex feature fusion mechanism, the proposed model can achieve the semantic correspondence between the line drawings and the reference maps. Our method can produce high-quality colored results with accurate semantic correspondence, especially in the regions of anime character faces.

3.4.2 Phase 2 Training

After training ϵ_θ, diffusion models inference through the learned reverse process. Since the result distribution of forward process $p(I_{\text{noise}}^{(T)})$ approximates a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, the sampling process starts from pure Gaussian noise, followed by T rounds of denoising. On one hand, training diffusion models often require a large batch size and long iteration rounds with large computing consumption. On the other hand, we want to strike a balance between the diversity and accuracy of generated results. After phase 1 training, the model should have already good denoising ability, we then introduce the image reconstruction guidance to improve the generation ability of AnimeDiffusion in phase 2 training.

We perform a small number of steps S of refinement in phase 2 based on the phase 1 trained model, as shown in Algorithm 1. Since the great diversity of the generated results of the diffusion model, as the training iterations grow, the quality of generated images is affected less by the input noise and more by the guidance condition. Therefore, we input the noise generated according to the reference image instead of random noise in the phase 2 training and inference process of the model. According to Eq. (6), I_{gt} is estimated as

$$\widetilde{I}_{\text{gt}} = \frac{1}{\sqrt{\bar{\alpha}_t}}(I_{\text{noise}}^{(t)} - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(I_l, I_r, I_{\text{noise}}^{(t)}, t)). \quad (8)$$

The mean value of reverse process $p_\theta(I_{\text{noise}}^{(t-1)}|I_{\text{noise}}^{(t)}, I_l, I_r)$ is parameterized as

$$\widetilde{\mu}_\theta(I_{\text{noise}}^{(t)}, t) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \cdot \widetilde{I}_{\text{gt}} + \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\bar{\alpha}_t} \cdot I_{\text{noise}}^{(t)}. \quad (9)$$

With the estimation of $\widetilde{\mu}_\theta(I_{\text{noise}}^{(t-1)}, t)$, each iteration of reverse process is

$$I_{\text{noise}}^{(t-1)} = \widetilde{\mu}_\theta(I_{\text{noise}}^{(t)}, t) + \sigma_t^2\epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (10)$$

where σ_t is the sampling variance with $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.

The time and memory consumption by directly adopting the reverse process of DDPM is vast, so we use DDIM [34] as our denoising function, which is an alternative non-Markov chain denoising process with different sampling process

$$I_{\text{noise}}^{(t'-\Delta t)} = \sqrt{\bar{\alpha}_{t'-\Delta t}}\widetilde{I}_{\text{gt}} + \sqrt{1-\bar{\alpha}_{t'-\Delta t} - \eta\sigma_{t'}^2}\epsilon_\theta + \eta\sigma_{t'}\epsilon. \quad (11)$$

DDIM obtains a sub-sequence of time $[0, T)$, where t' is sampled time sequence [34], Δt is the interval in the sampling sequence t', η is a hyper-parameter that controls whether noise is added during the reverse process. If η is set to 0, the process of image generation is deterministic.

Since both the forward and reverse processes of DDPM are random, the colorization results are different for the same sample. DDIM provides a deterministic reverse sampling strategy, but due to the different initial noise, there is no guarantee that the image reconstruction can be completed with the original image as the reference image. To fully utilize the image synthesis capability of diffusion models, we borrow the deterministic forward process from DiffusionCLIP [19] in our phase 2 training. According to

Algorithm 1 AnimeDiffusion Phase 2 Training

Input: ϵ_θ phase 1 trained noise prediction model,
$\mathcal{I}$ training set of anime face images,
S number of sampling steps.
Output: $\hat{\epsilon}_\theta$ refined noise prediction model.

Initialize list of forward noisy images $\hat{\mathcal{I}}$;
Sampling a sub-sequence $\mathcal{T}' \in [0, T)$ of length $|\mathcal{T}'| = S$;
for I_{gt} **in** $\mathcal{I}$ **do**
  for t' **in** $\mathcal{T}'$ **do**
    Calculate $\widetilde{\epsilon} \leftarrow \epsilon_\theta(I_l, I_r, I_{\text{noise}}^{t'}, t')$;
    Predict the ground truth $\widetilde{I}_{\text{gt}}(I_{\text{noise}}^{t'}, t')$;
    Forward step $I_{\text{noise}}^{(t'+\Delta t)} \leftarrow \sqrt{\bar{\alpha}_{t'+\Delta t}}\widetilde{I}_{\text{gt}} + \sqrt{1-\bar{\alpha}_{t'+\Delta t}}\widetilde{\epsilon}$;
  end for
  Update $\hat{\mathcal{I}}$;
end for
for $I_{\text{noise}}^{\mathcal{T}'}$ **in** $\hat{\mathcal{I}}$ **do**
  for t' **in** `reverse`($\mathcal{T}'$) **do**
    Calculate $\widetilde{\epsilon} \leftarrow \epsilon_\theta(I_l, I_r, I_{\text{noise}}^{t'}, t')$;
    Predict the ground truth $\widetilde{I}_{\text{gt}}(I_{\text{noise}}^{t'}, t')$;
    Reverse step $I_{\text{noise}}^{(t'-\Delta t)} \leftarrow \sqrt{\bar{\alpha}_{t'-\Delta t}}\widetilde{I}_{\text{gt}} + \sqrt{1-\bar{\alpha}_{t'-\Delta t}}\widetilde{\epsilon}$;
  end for
  Update reconstruction loss L_{rec};
  Gradient step $\nabla_{\hat{\epsilon}_\theta} L_{\text{rec}}$;
end for

Eq. (8) and Eq. (11), DDIM is considered as an Euler method to solve an ordinary differential equation (ODE)

$$\mathrm{d}\frac{I_{\text{noise}}^{(t)}}{\sqrt{\bar{\alpha}_t}} = \epsilon_\theta \cdot \mathrm{d}\sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}. \tag{12}$$

The above ODE holds in a finite number of steps, and to obtain an accelerated forward process, we use the same sampling time series t' as the reverse process of DDIM. The recursive relation from $I_{\text{noise}}^{(t')}$ to $I_{\text{noise}}^{(t'+\Delta t)}$ is simplified as

$$I_{\text{noise}}^{(t'+\Delta t)} = \sqrt{\bar{\alpha}_{t'+\Delta t}}\widetilde{I}_{\text{gt}}(I_{\text{noise}}^{(t')}) + \sqrt{1-\bar{\alpha}_{t'+\Delta t}}\epsilon_\theta, \tag{13}$$

$\widetilde{I}_{\text{gt}}(I_{\text{noise}}^{(t')})$ represents the ground truth estimation function of Eq. (8). Accordingly, to obtain the deterministic reverse process, we apply η as 0 in the reverse process of DDIM.

After S iterations, the denoising function finally generates the colored image I_{gen}. We calculate the mean square error (MSE) between I_{gen} and I_{gt}, in order to constrain the generated image as close to the original image as possible in pixel level. It is worth noting that the dotted arrow in Fig. 2 represents reverse gradient propagation to update the parameters of U-Net during the phase 2 training. Our objective is

$$L_{\text{rec}} = \mathbb{E}_{I_{\text{gt}}\sim q(I_{\text{gt}}),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \parallel I_{\text{gen}} - I_{\text{gt}} \parallel_2 . \tag{14}$$

In summary, the hybrid training strategy can reduce computation consumption of training and inference significantly in comparison to typical training of diffusion method. The training objective of phase 1 is to obtain the solution of Eq. (12), the derivative of the path from the initial distribution to the target distribution at time step t. When applying Eq. (11) for inference, the distribution of the sub-sequence of time steps $\mathcal{T}'$ will affect the colorization results due to the deviation between the predicted derivatives ϵ_θ and the true path from time step t' to $t' - \Delta t$. Therefore, we fix the sub-sequence t' in the phase 2 training and select a small number of sampling steps to make our model save inference time. By refining the model trained in phase 1, so that the noise prediction at time step t' is closer to the direction pointing to the next time step $t' - \Delta t$, rather than accurately predicting the derivative.

4 Experimental Results

4.1 Dataset

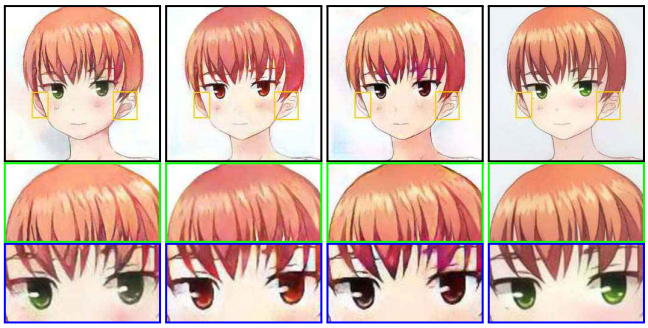
In this study, we focus on the anime face line drawing colorization task. To train AnimeDiffusion, a large dataset of high-resolution anime face images with paired line drawing and color data is required, and the image resolution should be high enough to adequately express the color and detail information of the face. Although some anime face work [41], [42] with datasets has been proposed recently, they are not designed for the purpose of colorization. We then build a benchmark dataset of anime face images. First, we collect anime character images from Danbooru2020 [20], which is a large-scale anime image database with 4.2m+ images. Next, we crop the face part of the character images according to our task requirements. After simple manual alignment and denoising operation, a dataset with 31,696 training data and 579 testing data of anime face images is obtained. Considering GPU memory restriction and model computational efficiency, all images are resized to 256×256 resolution. To generate paired line drawing images simulating artwork of artists, we use XDoG [39] to extract line drawings from colored anime images and set the parameters of XDoG algorithm with $\phi = 1\times10^9$ to keep a step transition at the border of lines in line drawings. We randomly set σ to be $0.3/0.4/0.5$ for different line widths, which generalizes AnimeDiffusion on various line styles to avoid overfitting. We set $p = 9, k = 3.5, \epsilon = 0.01$ in XDoG.

4.2 Implementation Details and Metrics

We devise our AnimeDiffusion via the PyTorch framework, and it is trained on 1 NVIDIA A100 GPU. The size of all input images is set as 256×256. For the diffusion hyper-parameters setting, we use a linear noise schedule of $(1e^{-6}, 1e^{-2})$ with T of 1000 time steps. We set a batch size of 32, and train the model for 300 epochs in phase 1, and we use a batch size of 4 to refine the model for 1 epoch in phase 2 training. On our device, the phase 1 training took 40 hours and the phase 2 training took 110 minutes. We apply the Adam optimizer with a learning rate of $1e^{-5}$ for both of the above stages.

Data augmentation: In practical scenarios, it is common to have large space discrepancies between the target line drawing and reference image in a colorization task. To enable AnimeDiffusion to learn accurate semantic correspondence during training and to avoid learning trivial solutions from pixel-aligned training data, we randomly deform I_{gt} by TPS transformation as the color references when loading the data for training. In other words, each batch of data will

(a) reference image (b) line drawing (c) original image



(d) Lee *et al.* [9] (e) Li *et al.* [10] (f) Cao *et al.* [11] (g) Our

Fig. 3. Detailed comparison of colorization results. Our model shows the best visual quality in these three relevant regions marked in different colors. We can colorize line drawings with accurate color consistency and semantic correspondence.

have different geometry distortions. To some extent, this is a data augmentation trick.

Evaluation Metrics: We mainly use three evaluation metrics for the quantitative comparison of AnimeDiffusion with other methods. The popular Fréchet Inception Distance (FID) is used to assess the generation ability of algorithms at the perceptual level. Besides measuring the perceptual credibility, we also adopt the Peak Signal-to-Noise Ratio (PSNR) and Multi-Scale Structural Similarity Index Measure (MS-SSIM) to evaluate the image reconstruction ability of algorithms in the pixel level. For quantitative evaluation, the higher the values of PSNR and MS-SSIM, the better the model represents, while the smaller the value of FID, the better the model is.

4.3 Quantitative Evaluations

We mainly compare AnimeDiffusion with three state-of-the-art GAN-based methods, including Lee *et al.* [9], Li *et al.* [10], Cao *et al.* [11]. We design two kinds of colorization tasks including self-reference reconstruction and random-reference colorization to analyze the colorization performance of AnimeDiffusion.

Self-Reference Reconstruction: In terms of reconstruction evaluation, the line drawing and reference image are paired. Ideally, the colorized output should be the same as the reference image. We directly use our paired testing data for self-reference reconstruction. In fact, during the training phase, AnimeDiffusion is trained to do image reconstruction; through this proxy task, the network learns the ability of colorization. For fairness, we train AnimeDiffusion and other three GAN-based methods using our training data. We then conduct a self-reference reconstruction task using the testing data to compute PSNR and MS-SSIM, respectively. The results are shown in Table 1 that AnimeDiffusion has the best image reconstruction performance, in terms of both PSNR and MS-SSIM scores.

TABLE 1
Quantitative Comparison with SOTA GAN-based Methods

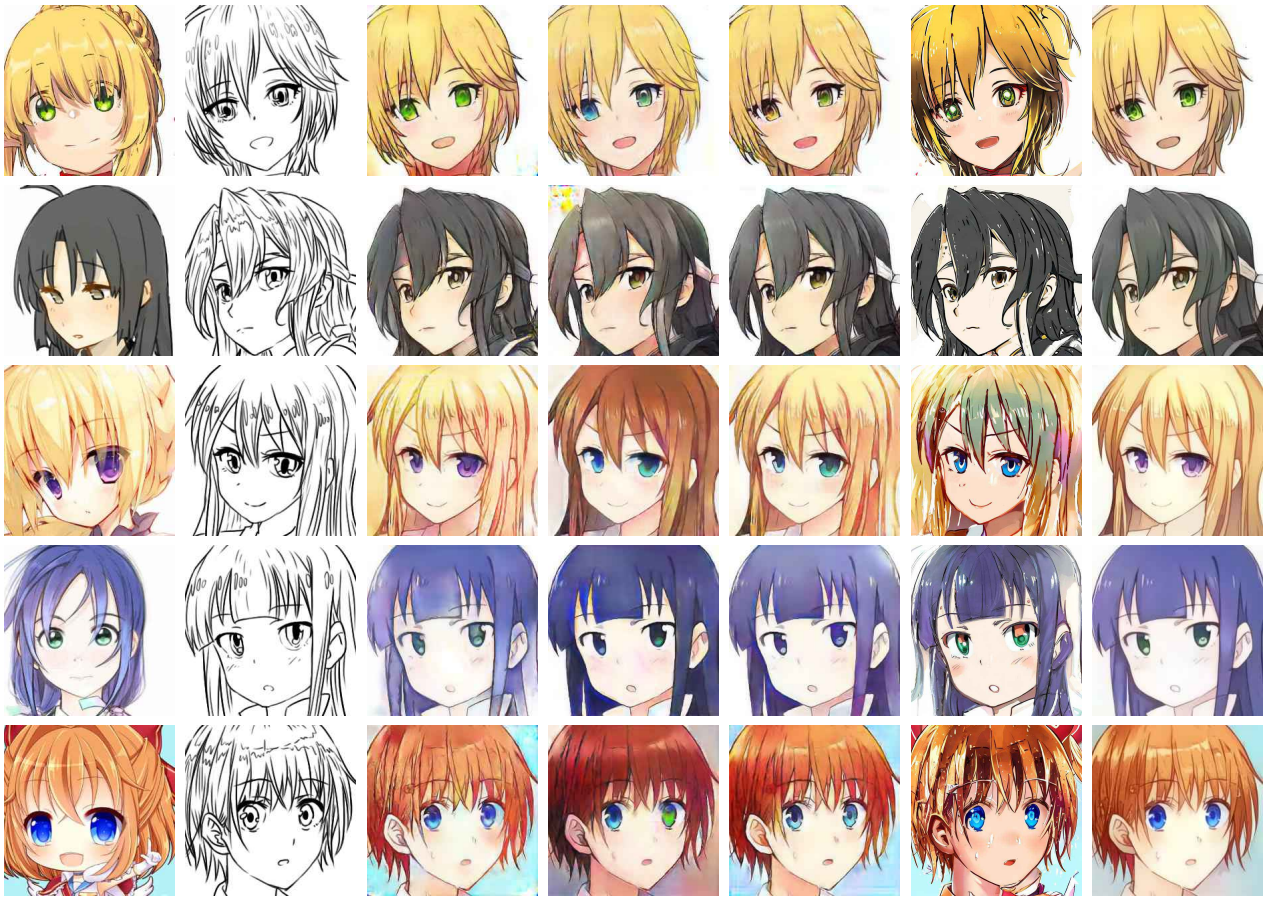
Method	PSNR↑	MS-SSIM↑	FID↓
Lee *et al.* [9]	23.8901	0.9224	57.19
Li *et al.* [10]	18.6347	0.8209	49.33
Cao *et al.* [11]	19.7746	0.8388	46.39
AnimeDiffusion	**25.4658**	**0.9596**	**44.19**

Random-Reference Colorization: We shuffle and unpair the line drawings and reference images in the testing data. Each line drawing is assigned with a random color reference image instead to perform random-reference colorization. The resulting 579 colorized results and 579 reference images are used to compute the FID score. A smaller FID indicates that the distribution of the colored results is closer to that of reference images and a better image generation ability of the model. As is shown in Table 1, AnimeDiffusion shows the best generation ability, in terms of FID, in comparison to the other three GAN-based methods.

4.4 Qualitative Evaluations

We also evaluate AnimeDiffusion qualitatively in comparison to the three state-of-the-art GAN-based methods [9], [10], [11]. For qualitative comparison, we set two cases of comparison, anime faces with *homochromatic pupils* and those with *heterochromatic pupils*. The latter case is more challenging, which demands a high level of semantic correspondence and precise local features. Fig. 3 shows a detailed comparison for the case of homochromatic pupils. The yellow region shows that AnimeDiffusion can recognize the ear semantic information from line drawing and inject the correct color the same as the face. The green region shows that AnimeDiffusion can generate the shiny hair as that of the original color image (Fig. 3(c)). The blue region indicates that AnimeDiffusion can accurately transfer the color information of the eyes from the reference image (Fig. 3(a)) into the line drawing (Fig. 3(b)). The other three methods all have flaws in these relevant areas. Since diffusion models are based on maximum likelihood theory with a more accurate estimation of the probability density than GAN-based methods do, the colored results are much more precise with little noise. Moreover, due to our training strategy, denoising and reconstruction tasks are separated during the training procedure, making the network training more stable with a better ability to capture detailed features.

More comparisons of homochromatic pupils are given in Fig. 4. In addition to GAN-based SOTA methods, we also compare to the diffusion model. Since we are the first attempt using diffusion model for reference-based colorization task, and other diffusion models for conditional image synthesis, e.g., [16], [44], are mainly guided by specially designed text prompts, while such text labels are not available or used in our task. We thus only make a comparison to the Prompt-Free model [8], which is an SD-based method tackling the exemplar-based image translation task. Given line drawings and reference images, AnimeDiffusion effectively generates colored results with accurate color and



(a) reference images (b) line drawings (c) Lee *et al.* [9] (d) Li *et al.* [10] (e) Cao *et al.* [11] (f) Xu *et al.* [8] (g) AnimeDiffusion

Fig. 4. Qualitative comparison of anime faces with homochromatic pupils. Given line drawings and reference images, AnimeDiffusion effectively generates colored results with accurate color and good semantic correspondence. The coloring results have a natural and smooth texture, precise colors with little noise, especially good in the eye colors, in which the sparkles in the eyes can be charmingly transferred. Compared with the other three GAN-based methods and one Stable Diffusion-based model, the image texture quality of AnimeDiffusion is far superior.

good semantic correspondence. The coloring results have a natural and smooth texture, and precise colors with little noise, especially for eye colors, in which the sparkles in the eyes can be charmingly transferred. In contrast, Lee *et al.* [9] generates results with color bleeding and incorrect semantic correspondence, the color texture is rough and the detail is blurry. The results of Li *et al.* [10] have obvious color flaws. Cao *et al.* [11] generates results with good image quality but the eyes are incorrectly colored. In general, the quality of images produced by GAN-based methods is not always stable and flaws randomly pop up. The result of [8] was an inference from the trained model with the default setting. As demonstrated in Fig. 4, although exemplar-based image translation is similar to the task of reference-based colorization and uses the same type of inputs, directly using this model to perform line drawing colorization cannot generate good results. Our results are of higher quality with more accurate colors and richer details.

For the case of heterochromatic pupils, Fig. 5 shows that AnimeDiffusion can generate results with accurate color in different pupils according to the reference images. Although Lee *et al.* [9] produces some results with heterochromatic pupils, the overall colorization quality is not high with color bleeding or flaws. The results of Li *et al.* [10] show globally distorted colors and some local errors in the eyes. Cao *et al.* [11] fails to color heterochromatic pupils, even though the overall image quality and semantic correspondence are good. Xu *et al.* [8] generates results with serious color bleeding. Heterochromatic pupils are fine-grained features in image space, all three SOTA GAN-based methods [9], [10], [11] and the diffusion-based method [8] cannot handle them well. Comparatively, our AnimeDiffusion has a better representation of the underlying data distribution, thus the pupils can be accurately colorized without introducing additional processing modules. Without using extra eye segmentation labels [4] or pupil position estimation [43], our method can generate better quality results with accurate color in pupils according to the reference image.

To verify the robustness of our method in terms of line styles and the effectiveness of intermediate representation of XDoG lines, we use three line drawings with different line styles downloaded from the internet for testing. Fig. 6 shows that our model has precise control over the line drawings of different styles, and the generated results have accurate colors with clear and complete line structures.

4.5 Ablation Study

We perform extensive ablation experiments to verify the effectiveness of the proposed hybrid training strategy. We

(a) reference images (b) line drawings (c) Lee *et al.* [9] (d) Li *et al.* [10] (e) Cao *et al.* [11] (f) Xu *et al.* [8] (g) AnimeDiffusion

Fig. 5. Qualitative comparison of anime faces with heterochromatic pupils. We are the first learning-based method that can generate results with accurate color in pupils according to the reference image with no extra eye segmentation label [4] or pupil position estimation [43]. Heterochromatic pupils are fine-grained features in the image space, all three SOTA GAN-based methods [9], [10], [11] and Stable Diffusion-based method [8] cannot handle them well in the task. Comparatively, AnimeDiffusion has a better representation of the underlying data distribution, thus the pupils can be accurately colorized without introducing additional processing modules.

find that the denoising model obtained by classifier-free guidance phase 1 training can generate images with high diversity, but this diversity also means that the colorization results are unstable since randomness is introduced by Gaussian noise. The classifier-free guidance phase 1 training aims to train a model with strong image denoising ability. In other words, the phase 1 trained model should already have the ability to capture the line structure and inject different colors into the corresponding areas, according to the reference image. The phase 2 training is only used to eliminate color gaps or align the color tone of the entire image. To validate this idea, we perform an image reconstruction test and a reference-based line drawing colorization test, respectively. We compare image quality with and without phase 2 training, and also being refined for 1 epoch or 10 epochs with a batch size of 4 in Fig. 7 and Fig. 8.

We perform the image reconstruction test using the ground-truth images as the color reference images, with results given in Fig. 7. The results in column (b) are obtained by performing 1000 steps of DDPM denoising (Eq. (7)) without any accelerated sampling strategy, using a phase 1 trained model. The colored results for models being refined based on loss Eq. (14) with a small number of time steps (S=10) from the phase 1 trained model for 1 epoch or 10 epochs are given in Fig. 7(c) and Fig. 7(d).

The reconstructed results without phase 2 refinement training have shown an aligned overall structure as the original image, but there is an obvious color difference between the two. After adding the image reconstruction loss in the phase 2 training, the coloring effect is significantly improved. The results of different refinement schemes are not obvious in terms of visual differences. For the reference-based line drawing colorization test, we show results in Fig. 8. Although the model without phase 2 training can distinguish regions of different colors, there is still a difference between the color tone of generated images and that of the reference images. Such color tone bias caused by random noises can be corrected and eliminated in phase 2 training.

We also compute PSNR, MS-SSIM, and FID metrics for quantitative comparison, and the results are shown in Table 2. After 10 epochs of refinement, AnimeDiffusion continues to gain in FID score but has little improvement in PSNR or MS-SSIM score. Furthermore, We conduct an ablation study to evaluate the effectiveness of the TPS transform. As shown in Table 2, although the value of PSNR, MS-SSIM, and FID is the best when the model is not equipped with TPS transform, the model just learns a trivial solution, i.e., directly outputs the reference images as

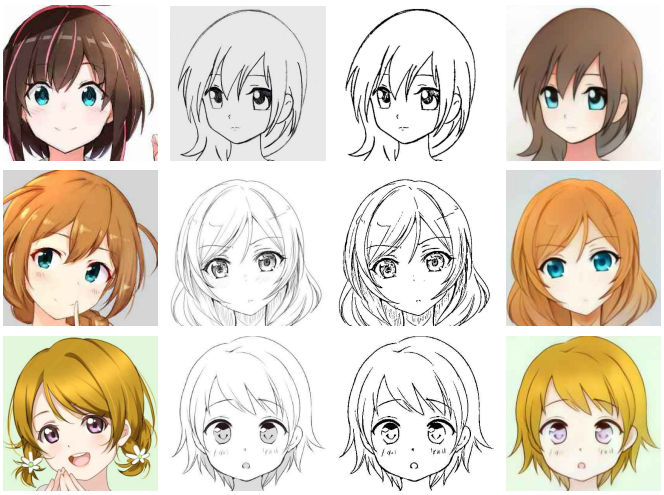


Fig. 6. Test results for line drawings of different line styles: (a) color reference images, (b) line drawings, (c) intermediate XDoG lines, (d) colorization results. Our model has a certain generalization performance for different styles of input line drawings while maintaining the line structure and accurately capturing colors including background and face colors in the reference image.

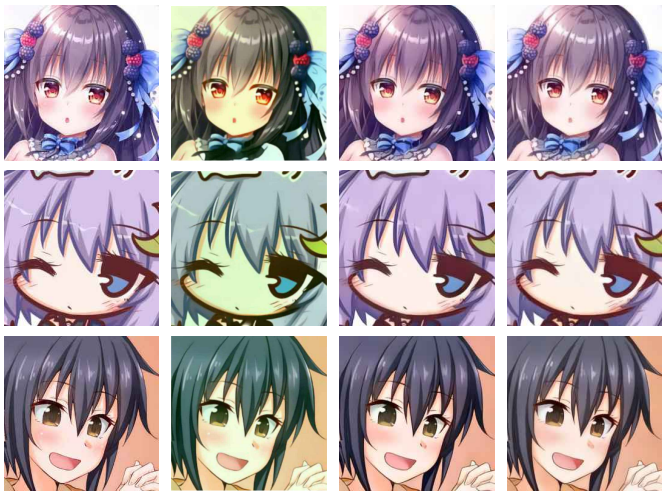


Fig. 7. Ablation study of image reconstruction: (a) ground-truth images, (b) results without phase 2 training, with 1000 steps of denoising, (c) results with refinement for 1 epoch, with 10 steps of denoising, (d) results with refinement for 10 epochs, with 10 steps of denoising. After adding the image reconstruction loss in the phase 2 training, the coloring effect is significantly improved. The results of different refinement schemes are not obvious in terms of visual differences.

TABLE 2
Ablation Study Results

AnimeDiffusion	PSNR↑	MS-SSIM↑	FID↓
Without TPS transform	33.2919	0.9909	9.2043
Without phase 2 training	12.4234	0.8079	55.1841
Refinement (1 epoch)	25.4658	0.9596	44.1876
Refinement (10 epochs)	**25.8992**	**0.9600**	**40.4392**

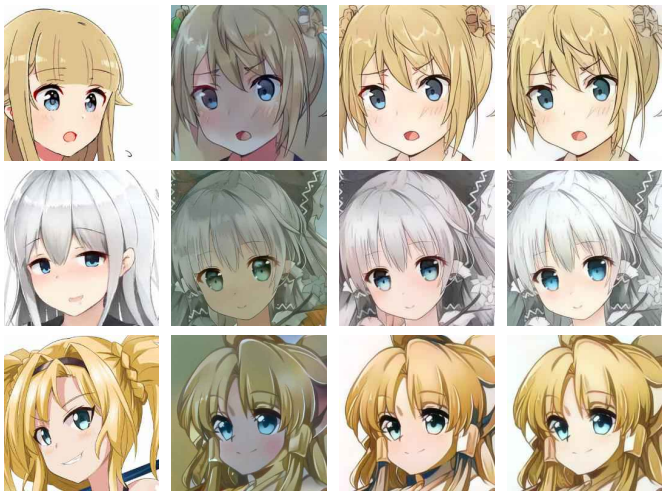


Fig. 8. Ablation study of reference-based line drawing colorization: (a) color reference images, (b) results without phase 2 training, with 1000 steps of denoising, (c) results with refinement for 1 epoch, with 10 steps of denoising, (d) results with refinement for 10 epochs, with 10 steps of denoising. Although the model without phase 2 training can distinguish regions of different colors, there is still a difference between the color tone of the generated images and that of the reference images, in which the colored images are dimmer.

the final colorization results without incorporating any semantic information. It illustrates that introducing distorted reference images can benefit the model in learning semantic information during the training stage.

In fact, refining the model for 1 epoch is sufficient for good colorization results. This validates that the phase 2 training is mainly used to fix color bias because the model trained in phase 1 already has very good image generation ability. The proposed hybrid training strategy allows effective model learning and saves training time cost. In Fig. 9, we visualize the loss curve for 400 epochs of the phase 1 training. It can be clearly seen that if purely relying on phase 1 training, the model is already converged after 300 epochs where the loss value from epoch 300 to 400 is kept at minimal. However, the color tone differences still exist due to different random noises in Fig. 7(b) and Fig. 8(b). Our proposed hybrid training strategy facilitates efficient model learning while simultaneously reducing the time expenditure associated with training.

4.6 User Study

It is challenging to evaluate the visual quality of generated images, in particular for line drawing colorization. A user study was conducted that subjects were shown with colorization results obtained by different colorization methods in comparison to color reference images, line drawings, and original color images. Two visual dimensions were evaluated, including color consistency and semantic correspondence. These evaluation criteria are very important for reference-based line drawing colorization tasks. Color consistency indicates whether the color information in the reference image is accurately extracted. Semantic correspondence means whether the color from the reference image is injected into the right place of the line drawing. Participants were asked to rate each of the two evaluation dimensions in Likert scale of 1-5. By comparing the coloring results with reference images and line drawings, human subjects

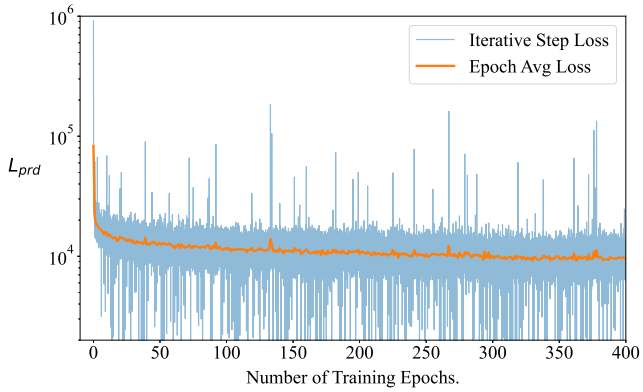


Fig. 9. Loss value in phase 1 training: Light blue represents the loss per iterative step and orange represents the average loss per epoch.

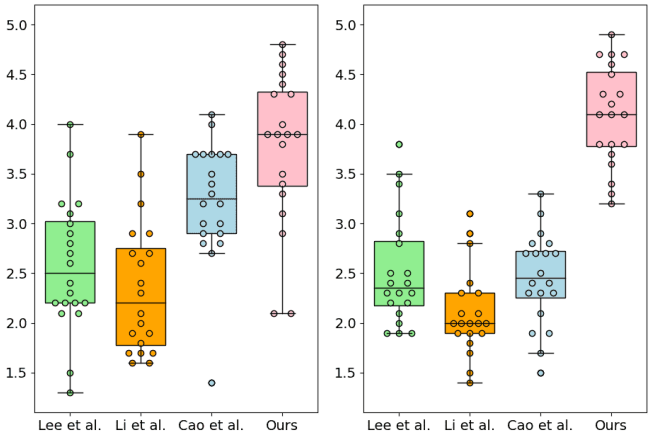


Fig. 10. Overall user study results from 20 participants. We set two visual evaluation dimensions for participants to rate on a five-point scale: (a) for color consistency and (b) for semantic correspondence. Each dot represents the score given by one participant. It indicates that our AnimeDiffusion achieves the best results by human assessment.

can visually judge the coloring effects of different methods. As shown in Fig. 10, in both dimensions of color consistency and semantic correspondence, our method outperforms the other three GAN-based methods, in terms of median and expectation. Through qualitative and quantitative analysis and user study, it indicates the superiority of our model for reference-based anime face line drawing colorization.

5 Application

5.1 User Interface and Inference Time

As is shown in Fig. 11, a user interface is developed for users to perform line drawing colorization using our AnimeDiffusion. Users only need to provide the line drawing and the reference image to AnimeDiffusion, and then the system generates colored results automatically without any human intervention. The time elapsed for the coloring process is printed at the bottom of the interface. With the hybrid training strategy and DDIM acceleration sampling strategy, our method generates high-quality colorization results efficiently. The time consumption for coloring a line drawing is within 0.2s on a machine equipped with

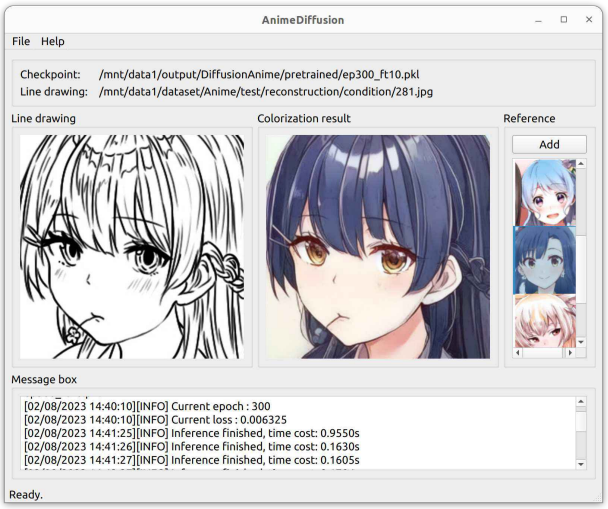


Fig. 11. User interface of AnimeDiffusion for colorizing anime face line drawings. Users only need to provide line drawings and reference drawings, and then complete the one-click operation of colorization. The time of the operation is printed at the bottom of the interface. The colorization time consumption for a line drawing can be controlled within 0.2s on a machine equipped with RTX 4090, excluding the initialization period.

RTX 4090, excluding the initialization period. Our end-to-end AnimeDiffusion model can be directly integrated into the practical colorization pipeline in the animation creation industry. In contrast to other diffusion methods [16], [35], our method can accurately edit face line drawings according to the given reference images.

5.2 Famous Anime Character Recolorization

Although we aim to train AnimeDiffusion to perform single line drawing colorization, the proposed method can also be used to recolorize a series of images or even consecutive video frames of the same anime character. In the animation creation industry, sometimes during the creation process, the same character appears in different colors in different images, and the proposed method is an efficient tool to standardize the character's colors based on one reference image. To do so, we first convert the original colored images to line drawings using XDoG extractor, then input these line drawings to AnimeDiffusion to generate colored results. Fig. 12 shows the recolorization results, in which for each anime character, the top row shows original colored images and the bottom row shows the recolorized results. As shown, according to one reference image, different images of the same character can be recolorized to a consistent color style.

5.3 Original Anime Character Colorization

We collaborate with professional artists and use AnimeDiffusion to colorize original line drawings of new character. The AnimeDiffusion has learned semantic information of anime character face, the model can generalize well to other hand-drawn characters, as demonstrated in Fig. 13. AnimeDiffusion can greatly save artists' creation time and help them to complete the animation more efficiently.

Fig. 12. Famous anime character recolorization: for each anime character, the top row shows input content images, the left image is color reference, and the bottom row shows the recolorization results. As shown, according to one reference image, different images of the same character can be recolorized to a consistent color style. The two anime characters are Hoshizora Rin and Sonoda Umi of LoveLive.

Fig. 13. Illustration of original anime character colorization. The AnimeDiffusion has learned semantic information about anime character faces, the model can generalize well to other hand-drawn characters. Although the anime character has an exaggerated hairstyle and complex eye structure, our model can still recognize and colorize them accurately. It indicates that our AnimeDiffusion has good generalization performance.

5.4 Fashion Sketch Colorization

We also extend AnimeDiffusion to colorize hand-drawn fashion design sketches into full-color fashion illustrations. Fashion illustration is similar to animation that designers first create designs in line drawings (fashion sketches) and then fill them with colors (fashion illustrations). We train AnimeDiffusion again on a proprietary dataset of fashion illustrations with a few hundred data. We keep the structure of the model and the training method the same. As shown in Fig. 14, given one color illustration reference and one line drawing design sketch, AnimeDiffusion can colorize the hand-sketch as a colored fashion illustration with accurate semantic correspondence of the face as well as the clothing and the body. This shows that our model can also handle other types of data and expand the colorization region from the face to the body if enough high-quality training data is given.

Fig. 14. Fashion sketch colorization: (a) color references, (b) input design sketches, and (c) colorization results. Given one color illustration reference and one line drawing design sketch, AnimeDiffusion can perform accurate semantic colorization in the face as well as the clothing and the body. It demonstrates that our model can generalize well to the colorization for the full body.

Fig. 15. Example failure case 1: (a) color reference, (b) colorization result, and (c) ground truth. When the input line drawings have semantic information that does not have sufficient training data in the dataset, the coloring result will be affected. See the Chibi cartoon example, teeth are not correctly identified by our model, and the color of the teeth is not accurately reflected in the result.

6 Conclusion and Future Work

In this paper, AnimeDiffusion – the first diffusion model tailored for reference-based anime face line drawing colorization is developed. We design a new hybrid training strategy for our diffusion model, separating the image denoising task and the image reconstruction task. Our model no longer follows the diffusion equation in ODE form for any sampled

(a) (b) (c)

Fig. 16. Example failure case 2: (a) color reference, (b) line drawing, and (c) colorization result. It shows that our model cannot handle abstract minimalist line styles and non-closed lines.

time series, but simulates the generation process from pure noise to colored anime in a specific few time steps. This approach not only offers a solution to impose high-level control on the diffusion model, through our specific task that best balances the fidelity and diversity for conditional image synthesis, but also improves the efficiency of both model training and inference. Extensive experiments and a user study have shown that AnimeDiffusion outperforms, both qualitatively and quantitatively, other state-of-the-art GAN-based methods and another diffusion-based method. The coloring results are of high image quality and with precise semantic color information. There are two limitations to the current method. Our model uses paired data for training, allowing it to learn semantic correspondence automatically. However, when the input line drawings with semantic information only have limited training data in the dataset, the coloring result will be affected, see Chibi cartoon example in Fig. 15. On the other hand, when the line style is very abstract and there are non-closed areas, the coloring result has a color bleeding problem, as shown in Fig. 16. Since the computation power limitation of our training device, we cannot handle very high-resolution (over 256×256) image colorization. In the future, we will work on line drawing colorization with multi-modal inputs, such as combining text information and reference images together for coloring line drawings of richer details.

ACKNOWLEDGMENTS

The authors would like to thank the editors and anonymous reviewers for their insightful comments and suggestions.

REFERENCES

[1] Y. Qu, T.-T. Wong, and P.-A. Heng, "Manga colorization," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 1214–1220, 2006.
[2] D. Sýkora, J. Dingliana, and S. Collins, "LazyBrush: Flexible painting tool for hand-drawn cartoons," in *Computer Graphics Forum*, vol. 28, no. 2, 2009, pp. 599–608.
[3] D. Varga, C. A. Szabó, and T. Szirányi, "Automatic cartoon colorization based on convolutional neural network," in *International Workshop on Content-Based Multimedia Indexing*, 2017, pp. 28:1–28:6.
[4] S.-Y. Chen, J.-Q. Zhang, L. Gao, Y. He, S. Xia, M. Shi, and F.-L. Zhang, "Active colorization for cartoon line drawings," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 2, pp. 1198–1208, 2022.
[5] X. Liu, W. Wu, C. Li, Y. Li, and H. Wu, "Reference-guided structure-aware deep sketch colorization for cartoons," *Computational Visual Media*, vol. 8, pp. 135–148, 2022.
[6] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5142–5152.
[7] X. Zhou, B. Zhang, T. Zhang, P. Zhang, J. Bao, D. Chen, Z. Zhang, and F. Wen, "CoCosNet v2: Full-resolution correspondence learning for image translation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 460–11 470.
[8] X. Xu, J. Guo, Z. Wang, G. Huang, I. Essa, and H. Shi, "Prompt-free diffusion: Taking "text" out of text-to-image diffusion models," *arXiv preprint arXiv:2305.16223*, pp. 1–12, 2023.
[9] J. Lee, E. Kim, Y. Lee, D. Kim, J. Chang, and J. Choo, "Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5800–5809.
[10] Z. Li, Z. Geng, Z. Kang, W. Chen, and Y. Yang, "Eliminating gradient conflict in reference-based line-art colorization," in *European Conference on Computer Vision*, 2022, pp. 579–596.
[11] Y. Cao, H. Tian, and P. Y. Mok, "Attention-aware anime line drawing colorization," in *IEEE International Conference on Multimedia and Expo*, 2023, pp. 1637–1642.
[12] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*, vol. 37, 2015, pp. 2256–2265.
[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 674–10 685.
[14] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.
[15] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *Journal of Machine Learning Research*, vol. 23, no. 47, pp. 1–33, 2022.
[16] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *IEEE International Conference on Computer Vision*, 2023, pp. 3836–3847.
[17] T. Xiao, S. Liu, S. De Mello, Z. Yu, J. Kautz, and M.-H. Yang, "Learning contrastive representation for semantic correspondence," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1293–1309, 2022.
[18] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "SDEdit: Guided image synthesis and editing with stochastic differential equations," in *International Conference on Learning Representations*, 2021, pp. 1–14.
[19] G. Kim, T. Kwon, and J. C. Ye, "DiffusionCLIP: Text-guided diffusion models for robust image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2416–2425.
[20] Anonymous, D. Community, and G. Branwen, "Danbooru2020: A large-scale crowdsourced and tagged anime illustration dataset," January 2021. [Online]. Available: https://www.gwern.net/Danbooru2020
[21] X. Dong, W. Li, X. Hu, X. Wang, and Y. Wang, "A colorization framework for monochrome-color dual-lens systems using a deep convolutional network," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 3, pp. 1469–1485, 2022.
[22] Y. Xiao, J. Wu, J. Zhang, P. Zhou, Y. Zheng, C.-S. Leung, and L. Kavan, "Interactive deep colorization and its application for image compression," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 3, pp. 1557–1572, 2022.
[23] F. Fang, T. Wang, T. Zeng, and G. Zhang, "A superpixel-based variational model for image colorization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 10, pp. 2931–2943, 2020.
[24] M. Xia, W. Hu, T.-T. Wong, and J. Wang, "Disentangled image colorization via global anchors," *ACM Transactions on Graphics*, vol. 41, no. 6, pp. 204:1–204:13, 2022.
[25] Y. Ci, X. Ma, Z. Wang, H. Li, and Z. Luo, "User-guided deep anime line art colorization with conditional adversarial networks," in *ACM International Conference on Multimedia*, 2018, pp. 1536–1544.
[26] L. Zhang, C. Li, T.-T. Wong, Y. Ji, and C. Liu, "Two-stage sketch colorization," *ACM Transactions on Graphics*, vol. 37, no. 6, pp. 261:1–261:14, 2018.
[27] H. Kim, H. Y. Jhoo, E. Park, and S. Yoo, "Tag2Pix: Line art colorization using text tag with SECat and changing loss," in *IEEE International Conference on Computer Vision*, 2019, pp. 9055–9064.
[28] C. Zou, H. Mo, C. Gao, R. Du, and H. Fu, "Language-based colorization of scene sketches," *ACM Transactions on Graphics*, vol. 38, no. 6, pp. 233:1–233:16, 2019.

[29] X. Li, B. Zhang, J. Liao, and P. V. Sander, "Deep sketch-guided cartoon video inbetweening," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 8, pp. 2938–2952, 2022.
[30] M. Shi, J.-Q. Zhang, S.-Y. Chen, L. Gao, Y.-K. Lai, and F.-L. Zhang, "Reference-based deep line art video colorization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 6, pp. 2965–2979, 2023.
[31] S.-Y. Chen, F.-L. Liu, Y.-K. Lai, P. L. Rosin, C. Li, H. Fu, and L. Gao, "DeepFaceEditing: Deep face generation and editing with disentangled geometry and appearance control," *ACM Transactions on Graphics*, vol. 40, no. 4, pp. 90:1–90:15, 2021.
[32] B. Yang, X. Chen, C. Wang, C. Zhang, Z. Chen, and X. Sun, "Semantics-preserving sketch embedding for face generation," *IEEE Transactions on Multimedia*, vol. 25, pp. 8657–8671, 2023.
[33] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
[34] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021, pp. 1–20.
[35] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH*, 2022, pp. 15:1–15:10.
[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, vol. 139, 2021, pp. 8748–8763.
[37] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," *arXiv preprint arXiv:2204.06125*, pp. 1–27, 2022.
[38] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, J. Ho, D. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.
[39] H. Winnemöller, J. E. Kyprianidis, and S. C. Olsen, "XDoG: An eXtended difference-of-Gaussians compendium including advanced image stylization," *Computers & Graphics*, vol. 36, no. 6, pp. 740–753, 2012.
[40] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567–585, 1989.
[41] Z. Huang, H. Xie, T. Fukusato, and K. Miyata, "AniFaceDrawing: Anime portrait exploration during your sketching," in *ACM SIGGRAPH*, 2023, pp. 14:1–14:11.
[42] Z. Li, Y. Xu, N. Zhao, Y. Zhou, Y. Liu, D. Lin, and S. He, "Parsing-conditioned anime translation: A new dataset and method," *ACM Transactions on Graphics*, vol. 42, no. 3, pp. 30:1–30:14, 2023.
[43] K. Akita, Y. Morimoto, and R. Tsuruno, "Colorization of line drawings with empty pupils," in *Computer Graphics Forum*, vol. 39, no. 7, 2020, pp. 601–610.
[44] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng, "MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," in *IEEE International Conference on Computer Vision*, 2023, pp. 22 560–22 570.

Yu Cao received the B.Eng. degree in communication engineering from the Qingdao Institute of Technology, Qingdao, China, in 2017, and the M.Eng. degree in communication and information system from the Xidian University, Xi'an, China, in 2020. He is currently pursuing the Ph.D. degree in fashion and textiles with The Hong Kong Polytechnic University, Hong Kong. His current research interests include line drawing colorization, AI for computer graphics, deep learning, and fashion colorization.

Xiangqiao Meng received the B.Eng. degree in coastal engineering and the M.Sc. degree in aerospace information technology both from the Zhejiang University, Hangzhou, China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree in computing with The Hong Kong Polytechnic University, Hong Kong. His current research interests include image colorization, diffusion models, and image synthesis.

P. Y. Mok (Member, IEEE) received the B.Eng. degree (Hons.) and the Ph.D. degrees in industrial and manufacturing systems engineering from The University of Hong Kong, in 1998 and 2002, respectively. She is currently an Associate Professor with The Hong Kong Polytechnic University, Hong Kong. Her current research interests include fashion 2D and 3D CAD, digital human modelling, cloth simulation, deep learning, sketch and pattern designs, computer graphics in fashion, and fashion design and synthesis.

Tong-Yee Lee (Senior Member, IEEE) received the Ph.D. degree in computer engineering from Washington State University, Pullman, in 1995. He is currently a Chair Professor with the Department of Computer Science and Information Engineering, National Cheng-Kung University (NCKU), Tainan, Taiwan. He leads the Computer Graphics Group, Visual System Laboratory, NCKU (http://graphics.csie.ncku.edu.tw). His current research interests include computer graphics, non-photorealistic rendering, medical visualization, virtual reality, and media resizing. He is a Senior Member of the IEEE and a Member of the ACM. He is an Associate Editor of the *IEEE Transactions on Visualization and Computer Graphics*.

Xueting Liu (Senior Member, IEEE) received the B.Eng. degree in computer science and technology from the Tsinghua University, Beijing, China, in 2009, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2014. She is currently an Assistant Professor with the School of Computing and Information Sciences, Caritas Institute of Higher Education, Hong Kong Her current research interests include computer graphics, computational art, and intelligent art.

Ping Li (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2013. He is currently an Assistant Professor with the Department of Computing and an Assistant Professor with the School of Design, The Hong Kong Polytechnic University, Hong Kong. He has published over 200 top-tier scholarly research articles (e.g., TVCG, TPAMI, TIP, TNNLS, TMI, TMM, TCSVT, TCYB, TBME, TSMC, TII, AAAI, CVPR, ICCV, NeurIPS), pioneered several new research directions, and made a series of landmark contributions in his areas. He has an excellent research project reported by the *ACM TechNews*, which only reports the top breakthrough news in computer science worldwide. More importantly, however, many of his research outcomes have strong impacts to research fields, addressing societal needs and contributed tremendously to the people concerned. His current research interests include image/video stylization, colorization, artistic rendering and synthesis, realism in non-photorealistic rendering, computational art, and creative media.