IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 November 2023, accepted 20 December 2023, date of publication 26 December 2023,
date of current version 22 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347551

RESEARCH ARTICLE

Refining Line Art From Stroke Style Disentanglement With Diffusion Models

FANGLU XIE, MOTOHIRO TAKAGI, (Member, IEEE), HITOSHI SESHIMO, AND YUSHI AONO
NTT Human Informatics Laboratories, Kanagawa 239-0847, Japan
Corresponding author: Fanglu Xie (fanglu.xie@ntt.com)

ABSTRACT A beginner who wants to create illustrations has difficulty improving his/her ability without expert advice. Especially in the initial steps, line drawings are critical but hard to evaluate because there are many assessment points, such as shape, variation in thickness, stroke fluency, and shadow expression. Moreover, there is no well-summarized line art dataset based on expert knowledge to support skill refinement. Furthermore, the evaluation criterion is always subjective. To solve this problem, we custom-build systematized line artworks formed by cataloged stroke styles and propose a machine learning method that can automatically give clues to refining the artworks. We request 10 professional-level artists to create line art in six patterns; the stroke styles of the images are systematically summarized. Using this specific dataset, we train an auxiliary classifier to identify and remove features of those patterns to refine all line artwork commonly. We also implement an enhancement step that uses diffusion models to add more informative details to the generated results. The proposed method can automatically identify where strokes are needed to change and generate high-quality refined versions. Our method performs better than the previous method regarding L2, lpips, and SSIM scores while giving specialized clues to different stroke styles.

INDEX TERMS Disentangled representation, image generation, line art refinement, denoising diffusion probabilistic models.

I. INTRODUCTION

The two main steps of illustration creation are line drawing and coloring. Illustration beginners always make mistakes in drawing lines and fail to realize where they went wrong. However, the evaluation of line drawings is complicated because there are too many assessment points, such as variation in thickness, stroke fluency, shadow expression, and drawing style. Furthermore, the evaluation criterion is always subjective based on the artists' experience and preferences. If beginners are trying to enhance line drawings, they need advice from experts. However, it is impractical to expect that dedicated tutors can provide advice anytime and anywhere. Our solution is to create an automatic line refinement system based on expert knowledge that can give beginners specific advice and generate refined artwork, as a practical substitute for experts.

The associate editor coordinating the review of this manuscript and approving it for publication was Songwen Pei.

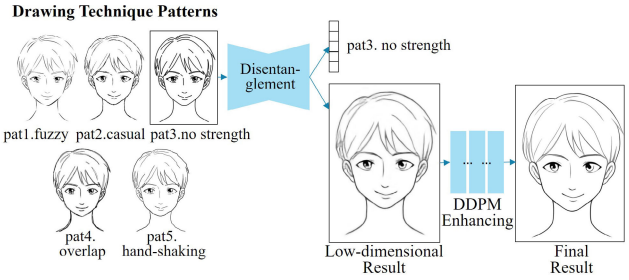


FIGURE 1. Example of proposed line art refinement. Features from line art patterns(*fuzzy, casual, overlap, no strength, hand-shaking*) can be distinguished and handled appropriately. Our method first distinguishes the stroke style and removes them in the latent space. Then the content features are converted into a low-dimensional refined result; the low-dimensional result is converted into a final version with clear lines through the denoising enhancing step.

An important issue is how to build a system that can summarize and apply expert knowledge. We analyze the composition of line art. Line artwork encompasses both content and style. The content determines where the lines

should be drawn, and it also reflects the artist's interpretation of the object's shape. In this paper, we refer to all these aspects collectively as "content." On the other hand, the style is reflected in the shape of the lines. The style also heavily depends on the artist, as it represents the characteristics of brush strokes and drawing techniques, such as how the line thickness changes and the lines connect. To make a system to provide professional and targeted advice, We consulted experts and referred to textbooks. Drawing insights from an art book [1], we identify five different stroke styles.

Many studies are closely related to our problem setting, such as research on image style transfer, sketch simplification, image disentanglement, etc. Traditional versatile image transfer strategies [2], [3] struggle to effectively capture and separate the distinct style attributes from underlying content. When these methods are applied to line artwork, they encounter difficulties in separating the complex line styles because of the unique slender characteristic of lines. There are other methods designed especially for the topic of line art refinement. Liu et al. [4] demonstrate that for successful style transfer in line art, the perceptual understanding of content should reach the level of perceiving centerlines(a proximate of the line topology). These methods indicate that the key to a successful line refinement is extracting correct content features. Sketch simplification [5], [6], [7] has the goal of cleaning line artwork automatically. However, these methods refine all artwork in a uniform and simplified style, and so lose a lot of significant details, especially variation in thickness. A key function of our proposal is to retain the line refinements that yield impactful artistic features.

To address the issues mentioned above, we propose a disentanglement strategy [8], [9], [10] that can extract the common features across different styles of line art and distinguish the style pattern at the same time. We then implement an autoregressive process based on a conditional denoising diffusion probabilistic model(DDPM) [11], [12], [13] to refine the line art while retaining impactful details. In this paper, the main contributions are as follows:

- Creation of a publicly available cataloged line art dataset using stroke styles.[1]
- Development of specialized models to enhance the disentanglement ability, allowing for the extraction of different style features and precise content features from line drawings. These models also provide style-specific suggestions.
- Generation of high-quality refined line art characterized by smooth and continuous lines with variations in their strength.

II. DATASET

Image datasets grouped into systematic patterns are very important to realize the disentanglement of stroke style features. Our team collaborated with 12 professional artists to create a diverse dataset. The first artist led the design of character images, while the second artist evaluated the artwork's quality and helped us choose the remaining 10 artists. The evaluator personally selected 10 artists to ensure that they were all professionals with unique artistic styles that covered a broad range of skills. Once selected, these 10 artists were responsible for rendering the characters in various stroke styles, based on the designed images. They are asked to create line artwork in six patterns, one is used as a sophisticated pattern (ground truth), and the other five are mediocre patterns with five kinds of stroke details. To ensure that the image information can be systematically disentangled into content and stroke style, we limited the target object to frontal face drawings in specified patterns. All of the artists drew the artworks of 10 characters using clear lines in their own style, which are taken as sophisticated data.

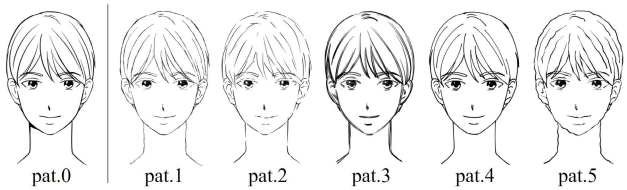


FIGURE 2. The target and five line artwork patterns with annotations: pat.0 ground truth(GT), pat.1 *fuzzy*, pat.2 *casual*, pat.3 *overlap*, pat.4 *no strength*, pat.5 *hand-shaking*.

TABLE 1. User study: 20 volunteers identified their favorite stroke style pattern.

pat.0 ground truth	pat.1 *fuzzy*	pat.2 *casual*	pat.3 *overlap*	pat.4 *no strength*	pat.5 *hand-shaking*
60%	10%	2%	11.5%	15%	1.5%

Our summarization is based on patterns mentioned in an art textbook [1]. We define mediocre patterns as fuzzy, casual, overlapping, lacking strength, and hand-shaking. Starting from the sophisticated types, all artists drew five kinds of line artwork with different stroke details, as shown in Fig.2.

- **pat.1 fuzzy**: carefully connecting short lines little by little makes the lines look weak and messy.
- **pat.2 casual**: trying to draw vigorously leads to a lot of broken bent lines that have uneven thickness and length.
- **pat.3 overlap**: lots of long lines stacked on top of each other which looks messy.
- **pat.4 no strength**: no variance in the line thickness gives the feeling of rigidity.
- **pat.5 hand-shaking**: drawing slowly and carefully yields soggy lines.

Based on the reference book, these five patterns cover the stroke styles seen most often in line sketch art. The expected refinement suggestions for pat.1 *fuzzy* (resp. pat.3 *overlap*) are to remove redundant short (resp. long) strokes that deviate from the correct line position. For pat.2 *casual*, the refinement is to reconnect the broken lines and set them in suitable positions. In the case of pat.4 *no strength*, the refinement is to enrich line thickness by adding or removing pixels of the lines at appropriate positions. For pat.5 *hand-shaking*, the network should smooth the curves into straight lines.

[1]Our dataset is available at https://github.com/ntthilab-generation/lineart_dataset

We think that relying solely on subjective evaluation criteria to create a dataset may make our makes experiments less credible. So, we conducted a user study with 20 volunteers to test the appeal of our sophisticated pattern designs, shown in Table.1. We experimented with different stroke styles and asked participants to rate their preferences to choose their favorite pattern. The results showed that 60% of respondents favored pattern.0, consistent with our established ground truth(GT). This result shows that the preference of the public is the same as our professional evaluator, which validates our request.

Overall, our dataset consists of 600 grayscale images, specifically 10 character head images in six patterns drawn by 10 artists. The image size is 1024×1024. The six patterns have an equal proportion in the data. We split the characters in the ratio of $7:3$ by artists for *training* : *testing*; yielding 420 training images and 180 test images.

III. RELATED WORK

We briefly review previous image refinement methods and analyze the characteristics of each method.

A. SKETCH SIMPLIFICATION WITH VECTORS

Several traditional methods simplify sketches by geometric processing [14], [15], [16], [17]; unfortunately, they accept the input of just stroke vectors, which means that they are unable to process complex real-world sketches. Igarashi et al. [14] simplify all sketched-type strokes by replacing each stroke group with a smooth curve, but this results in the loss of detail and poor-quality artwork. The methods of [15] simplify vector images by calculating the closure area but fail to handle raster images. Liu et al. [16] propose a vector graph simplification method based on the agglomerative generation of a graphic primitive through the use of a hard threshold; it is unable to well handle arbitrary sketches. Cole et al. [17] utilize depth and silhouette information obtained from 3D models to evaluate the significance of input strokes. For stroke evaluation, this method employs item buffers and priority buffers that determine line visibility and line density, respectively. The main problem with these methods is that they delete existing strokes to simplify the input sketches resulting in monotonous and uninteresting lines.

B. STYLE TRANSFER ON LINES

From an analysis of line artwork construction, we posit that a line drawing consists of two components: content and style. Some technologies are designed based on this structure. There are learning-based strategies for style transfer [2], [18], [19] targeted to colored arts. They use deep learning algorithms to extract general style and content and optimize an image to achieve a visually appealing result that recreates the original content in the desired style. Gatys et al. [18] find that the representations of content and style are separable in the convolutional neural network. Johnson et al. [19] introduce perceptual losses, blending high-level features with pixel-based losses for efficient style transfer. Xun and Serge [2] introduce Adaptive Instance Normalization (AdaIN); it applies dynamic adaptation of style image normalization to content images for flexible style transfer. However, when these methods are applied to line drawings, they encounter challenges in extracting the special style. The slender lines make separating meaningful line style from content difficult. There is a style transfer proposal [4] that targets line art but it demands extra input. Their proposal is efficient in terms of achieving accurate line style transfer with extra centerlines(a proximate of the line topology). Based on a discussion of this work, we elucidate that the key point in deep line art style transfer is to separate the detailed style features from the precise content. All the previous works demonstrate that for successful style transfer in line art, the perceptual understanding of content must reach the level of the perception of centerlines.

C. MULTI-DOMAIN DISENTANGLEMENT

Multi-domain disentanglement [8], [9], [10] refers to extracting independent factors across multiple domains and the domain-dependent factors that characterize each domain. This method achieves good performance in extracting domain-independent features. If we treat the patterns of line art as different domains, our work is similar to a multi-domain disentanglement task. The content independent of the patterns is the domain-independent factor across multiple patterns, and the style feature dependent on the patterns is the domain-dependent factor. The methods of [8] and [9] employ auxiliary classifiers to learn domain-independent and domain-dependent features through conditional adversarial training. Yu et al. [10] utilize this structure to bridge multi-modal translation and multi-domain translation. Inspired by this assumption, we utilize conditional adversarial training to extract detailed style features from different patterns and content features across different patterns. Unlike these prior methods, our proposal extracts detailed features based on a systematic definition of line patterns to give structural advice without recourse to extra centerlines.

D. ROUGH SKETCH SIMPLIFICATION

Sketch simplification is the research topic that reduces a complex and busy draft to line art. Machine learning methods for sketch simplification [5], [6], [7] can clean raster sketches. Simo-Serra et al. [5] propose an automatic approach that uses a fully convolutional network for simplifying sketches directly from raster images. Mastering sketching [6] offers improved multi-usability by employing generative adversarial networks (GANs) [20] with training based on wild datasets. The method of [7] achieves detailed line refinement through manual selection by the user. This method achieves high-quality simplification, but the correction is based on the user's judgment making it unsuitable for illustration beginners. In addition, all these methods strongly simplify the sketches, resulting in fewer impactful details than real line artwork. They treat all artworks the same without

regard for their different drawing styles, which may result in unexpected refinement when processing extremely sketchy strokes. Unlike sketch simplification methods, our proposal can extract detailed stroke style features from different style patterns and provide high-quality refinement strategies.

E. ENHANCING IMAGES WITH GENERATIVE MODELS

To generate enough details to represent effective characterization of the different patterns, we design an image enhancement step on a generative model to further modify the results.

Two generative models have gained popularity recently due to their ability to generate images. One is GAN [20]; it utilizes two networks(a generator and a discriminator) to compete and generate samples. GANs are known for producing high-quality and sharp samples, such as generating realistic images. Ledig et al. [21] propose a very deep ResNet [22] architecture using the concept of GANs [20] to form a perceptual loss function for photo-realistic super-resolution single images. Simo-Serra et al. [6] utilize supervised adversarial training with a paired dataset and an unsupervised dataset(free sketches) to achieve data augmentation. Motivated by these methods, we made an experiment [23] on utilizing the GAN structure to enhance blurred line art. However, GAN training is unstable and suffers from issues like mode collapse, where the generator focuses on generating a subset of samples. Furthermore, based on our experiments [23], the generated results lack varieties of textures when training on the limited dataset.

The denoising diffusion probabilistic model(DDPM) [11] is another generative model. It is an auto-regressive model and uses diffusion processes to generate samples. Due to its autoregressive generation process, *DDPM* can generate high-quality images with complex and varied textures, which is suitable for refining line art. Conditional DDPM [12], [13], [24] are proposed to transfer parts of reference images by diffusion models. Palette [13] is a unified framework based on a conditional DDPM model for image-to-image translation tasks (colorization, inpainting, uncropping, and JPEG restoration) without task-specific customization or optimization instability. Our work introduces a conditional DDPM model based on [13] to make refined sketches clearer and more similar to real line art.

IV. PROPOSED METHOD

In this paper, we present a two-step line art refinement method and evaluate the capability of each step. Figure 3 shows the overall architecture. The main idea of our framework is to identify the different patterns of stroke styles and separate them from the content features. Only the content features are utilized to achieve efficient line art refinement. Towards this end, we design the first feature extraction step(*FE* step) that generates initial images independent of patterns. Then, based on common results, the second step is a further enhancement step to create high-quality images based on the DDPM method (*DDPM* step). First, we introduce the feature extraction step in Section. IV-A. Then, we describe the knowledge of the diffusion model and detail the second step in Section. IV-B.

A. DISENTANGLEMENT ON STROKE STYLE FEATURES

To disentangle the stroke style features(pattern-dependent style) from the content, we design two specific encoders, a style encoder E_s and a content encoder E_c, which extract the style and content features, respectively. We use the layers of the vgg network [25] to compose the encoders. In addition, an auxiliary classifier, C, is used to identify the patterns. This classification has two roles: one is to give accurate classification results and specified drawing comments; and the other is to help separate the style features from the content features. Through an adversarial learning method with a GRL [8] layer, this structure helps to yield pattern-dependent styles and separate them from the content features, simultaneously. The low-dimensional style and content features are fed to generator G to output rough interim images. All networks, E_s, E_c, C and G, are trained at the same time by using the sum of losses given by Eq. 1. $\mathcal{L}_{class}$, $\mathcal{L}_{rec}$, $\mathcal{L}_{con}$ and $\mathcal{L}_{enh}$ are introduced in the following paragraphs.

$$\mathcal{L}^{FE}_{total} = \mathcal{L}_{class} + \mathcal{L}_{rec} + \mathcal{L}_{con} + \mathcal{L}_{enh} \tag{1}$$

We assume that the pattern-dependent style feature $\boldsymbol{f}^i_s$ is used for tuning the line style while the common drawing information is present in the content style $\boldsymbol{f}_c$. $\boldsymbol{I}i$ denotes the input line artwork with pattern i and $\boldsymbol{I}0$ is the target line artwork(ground truth). Style encoder E_s and content encoder E_c are designed to extract the pattern-dependent style $\boldsymbol{f}^i_s = E_s(\boldsymbol{I}i)$ and the content $\boldsymbol{f}_c = E_c(\boldsymbol{I}i)$, respectively. They all use the same structure as the vgg network. We can get the complete drawing feature $\boldsymbol{f}i$ from the sketch by combining style and content in an adaptive instance normalization process [2] way as

$$\boldsymbol{f}_i = \mathrm{AdaIN}(\boldsymbol{f}^i_s, \boldsymbol{f}_c) = \sigma(\boldsymbol{f}^i_s)\left(\frac{\boldsymbol{f}_c - \mu(\boldsymbol{f}_c)}{\sigma(\boldsymbol{f}_c)}\right) + \mu(\boldsymbol{f}^i_s) \tag{2}$$

We expect that pattern-dependent style features $\boldsymbol{f}^i_s$ can be separated from the common feature and then grouped by our designed patterns in the latent space. To achieve that, we add a gradient reversal layer (GRL) [8] to E_c, following $\boldsymbol{f}^{grl}_c = GRL(\boldsymbol{f}_c)$. The encoders are then trained to disentangle $\boldsymbol{f}^i_s$ from $\boldsymbol{f}_c$ by an adversarial classification training process that minimizes

$$\mathcal{L}_{class} = \sum_{i=1} li \log\{C(\mathrm{AdaIN}(\boldsymbol{f}^i_s, \boldsymbol{f}^{grl}_c))\} \tag{3}$$

Combined feature $\boldsymbol{f}_i$ is classified by the patterns. $\mathcal{L}_{class}$ measures the cross entropy between the (ground truth) pattern label l_i and $C(\boldsymbol{f}_i)$. While in the backpropagation step, the gradient values are back-propagated to the parameters of the style encoder (θ_s) as usual, while the reverse is performed for

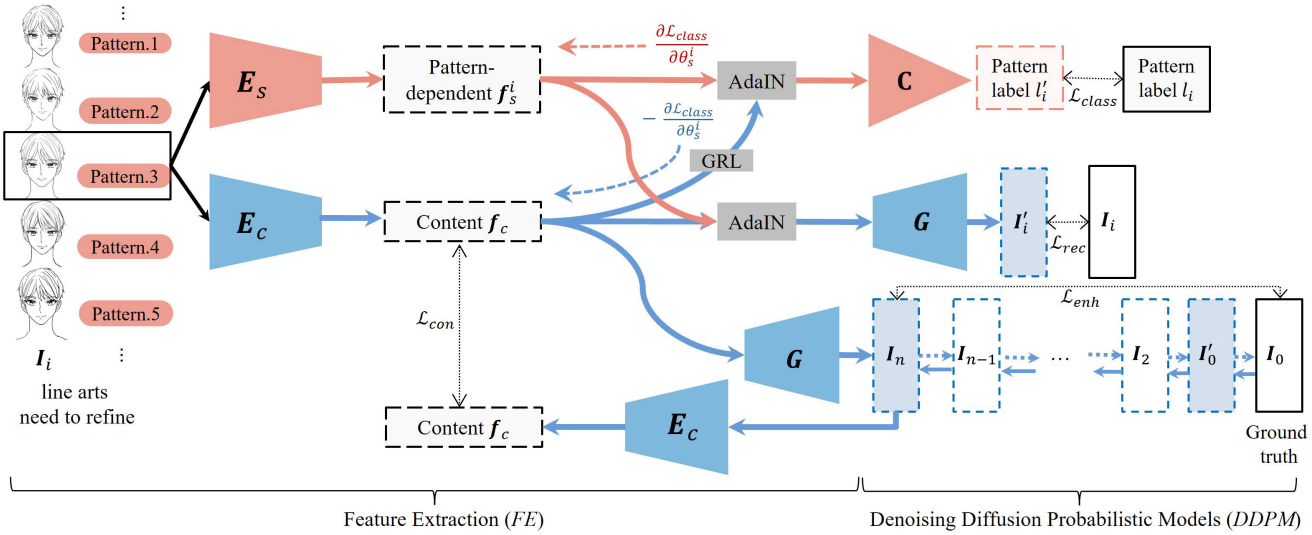


FIGURE 3. **Our two-step network architecture. In the *FE* step, style encoder E_s, content encoder E_c, generator *G*, pattern classifier *C*; all are trained simultaneously. In detail, given a sketch as input, the network is trained to distinguish and disentangle the pattern-dependent features from the content features. And only the content feature is pushed to the *G* to synthesize a roughly refined image. Then, in the *FE* + *DDPM* step, the rough results are transformed into high-detailed enhanced images through a conditional DDPM model.**

the content encoder (θ_c).

$$\theta_s \leftarrow \theta_s - \frac{\partial \mathcal{L}_{class}}{\partial \theta_s}, \quad \theta_c \leftarrow \theta_c - (-\frac{\partial \mathcal{L}_{class}}{\partial \theta_c}) \tag{4}$$

The reversed gradient values passed through the parameters of content encoder θ_c make it impossible to classify the pattern from content feature $\boldsymbol{f}_c$. That is, only the pattern-transparent component is pushed to $\boldsymbol{f}_c$, and, as a result, only the pattern-dependent component is expected to be retained in stroke style feature $\boldsymbol{f}_s^i$.

The generator needs to realize two generation processes. One reconstructs the original image from the input merged stroke style and content features, while the other generates a common result solely based on the input content features. To meet the reconstruction requirement, we design a function loss, $\mathcal{L}_{rec}$, that processes the input merged stroke style and content features of the original images. In addition, to ensure that the content is successfully extracted, we also set a loss function, $\mathcal{L}_{con}$, between the content feature of the input and the content feature of the generated results.

$$\mathcal{L}_{rec} = \sum_{i=1} ||G(\boldsymbol{f}_i) - \boldsymbol{I}_i||^2 \tag{5}$$

$$\mathcal{L}_{con} = \sum_{i=1} ||Ec(G(\boldsymbol{f}_c)) - \boldsymbol{f}_c||^2 \tag{6}$$

Simultaneously, we design another loss function, $\mathcal{L}_{enh}$, to refine images without pattern features by inputting the content features to generator *G*.

$$\mathcal{L}_{enh} = \sum_{i=1} ||\boldsymbol{I}' - \boldsymbol{I}_0||^2 = \sum_{i=1} ||G(\boldsymbol{f}_c) - \boldsymbol{I}_0||^2 \tag{7}$$

B. ENHANCING LINE ART WITH DIFFUSION MODEL

The *FE* step yields corrected lines that have accurate shapes but are blurred in style, see Fig. 1. We design the following steps using the conditional diffusion model to convert rough sketches into clean, high-quality images. Two feature encoders and the classifier load the parameters of the trained models after the *FE* step. The generated results from the *FE* step are adjusted by the *DDPM* training step by a U-net [26] structure. Diffusion models [11] convert samples from a standard Gaussian distribution into samples from an empirical data distribution through an iterative denoising process. Conditional diffusion models [13] make the denoising process conditional on an input signal. This takes the form of image-to-image processing following $p(\boldsymbol{I_0}|\boldsymbol{I'})$. Given a training output image $\boldsymbol{I'}$, we generate a noisy version $\tilde{\boldsymbol{I_0}} = \sqrt{\gamma}\boldsymbol{I_0} + \sqrt{1-\gamma}\epsilon$. We train the U-net $f\theta$ to denoise the noisy version with a noise level indicator γ, for which the loss is

$$\mathbb{E}(\boldsymbol{I'}, \boldsymbol{I_0})\mathbb{E}_{\epsilon\sim\mathcal{N}_{(0,I)}}\mathbb{E}\gamma \|f\theta(\boldsymbol{I'}, \sqrt{\gamma}\boldsymbol{I_0} + \sqrt{1-\gamma}\epsilon, \gamma) - \epsilon\|_1^1 \tag{8}$$

V. EXPERIMENTS

Our experiments evaluate performance in terms of disentanglement and image generation quality. To confirm successful disentanglement, we record the classification accuracy of the pattern-dependent feature and show that this feature can be separated from the content features. Furthermore, we compare the image generation results with state-of-the-art works. Image generation results are used to determine the quantitative error between results to the ground truth and several qualitative observations. Sketch simplification methods [6], [7] use pre-trained models while the other methods [2], [3], [10], [13] use the same dataset (our dataset) for training.

For the implement details, we use a machine comprising 2 NVIDIA RTX A6000 48GB. The first step,

TABLE 2. Average pattern classification accuracy from features.

Latent code	train	test
Ours ($AdaIN(f_s^i, f_c)$)	85.33	87.14
Ours (f_s^i)	68.00	69.14
Ours (f_c)	16.67	12.86

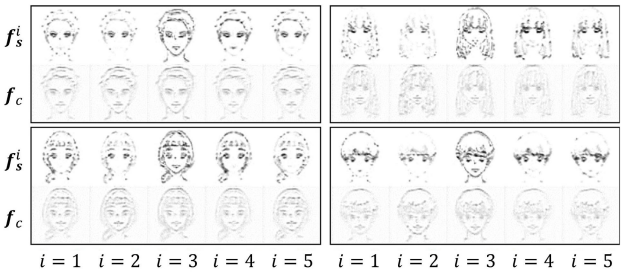


FIGURE 4. Comparison of a latent stroke style and content feature for pattern i.

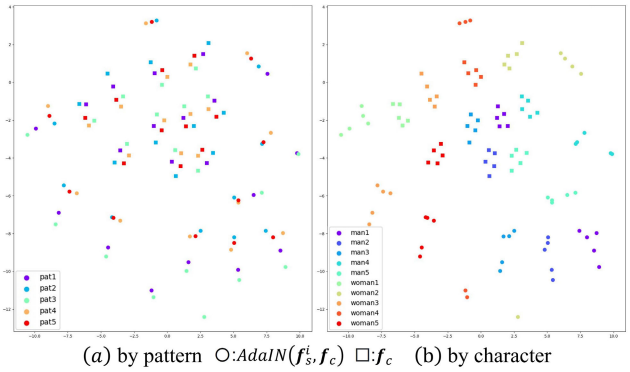


FIGURE 5. Feature visualization results of all images from one artist, labeled by patterns and characters. Each cluster covers all patterns and is marked by one character. f_c clusters of each character are tighter than those of $AdaIN(f_s^i, f_c)$.

FE, is trained for 1K epochs, while the second step, $FE + DDPM$, is trained for 3K epochs using the PyTorch framework. The number of diffusion steps is 2K during training, and the variances β_t are scaled linearly from 1×10^{-6} to 0.01. The code for the FE step utilizes the diffuseVAE [27] resource, but we modify the VAE structure by using 2 Vgg [25]. Additionally, we introduce an auxiliary classifier in 2 convolution layers and 3 linear layers, along with an AdaIN layer [2] and a GRL layer [8]. The code for the $DDPM$ step is derived from the Palette [13] sources.

A. DISENTANGLED REPRESENTATION

We conduct several evaluations to investigate the feature disentanglement performance of our method.

1) PATTERN CLASSIFICATION PERFORMANCE

We evaluate the average accuracy achieved in classifying the five patterns. First, we test the performance of classifier C; the results are shown in Table 2. The first row, $Ours(AdaIN(f_s^i, f_c))$, shows that training yields the classification accuracy of 85.33% for the combination of the stroke style features and the content features with the training data input. Moreover, this trained model also works for the test data (87.14%). The second row, $Ours(f_s^i)$, shows that stroke style details can also be identified in training and test data with accuracy of 68.0% and 69.14%, respectively, which shows that our network works when only stroke style features are input. If stroke style feature f_s^i is discarded, the accuracy drops to the level of chance (the third row $Ours(f_c)$) shows that content features cannot be identified.

These results show that the trained classifier can identify which type the stroke style feature belongs to. Furthermore, our method can distinguish the pattern-dependent (stroke style) components and pattern-transparent (content) components.

2) FEATURE VISUALIZATION

To analyze the information extracted from the content features, and the patterns and their relationships more intuitively, we output visual images of latent features and map the latent codes according to multiple labels.

We output t-SNE visualizations [28] of the extracted features in Fig. 6 based on the training dataset. The figures are color-coded to indicate the patterns, see Fig. 6(a), the clusters of f_s^i are consistent unlike those of f_c, see Fig. 6(c). In Fig. 6(b), when the pattern of stroke style is added to the f_c, the features points $(AdaIN(f_s^i, f_c))$ are separated. Furthermore, when f_c are labeled by the artists or characters (Fig. 6(e,f)), the latent code tends to concentrate with regard to the characters, but is smoothly distributed in terms of the artist. This situation shows that the content features are dependent on both the character and the artist. These results show that stroke style feature f_s^i can be disentangled successfully from content features f_c and separated successfully by the patterns.

3) DISPERSION ANALYSIS

To prove that our proposal extracts the same content information from different mediocre sketches, we further analyze the dispersion of latent codes by visual output and latent mapping.

We show visual images of latent features in Fig. 6 to analyze the relationship between stroke style features f_s^i and content features f_c. For example, when the input one-channel image size is 1024×1024, the latent codes are 64×64 blocks multiple by 512 channels. As sampled results(e.g. the 512-th channel) are shown in Fig. 4, the input images are the same item but drawn in patterns from pat.1 to pat.5. The first row are the visual images of f_s^i and the second row shows the visual images of f_c. We can see that the f_c results have a slender shape that characterizes the common drawing position across different patterns of stroke style. Moreover, although they are drawn using different stroke style patterns, they are close to each other. In contrast, f_s^i results are different for different patterns. It can be concluded that similar content features can be extracted from different mediocre input images, and the specified different patterns of stroke style can be extracted, which means that the patterns

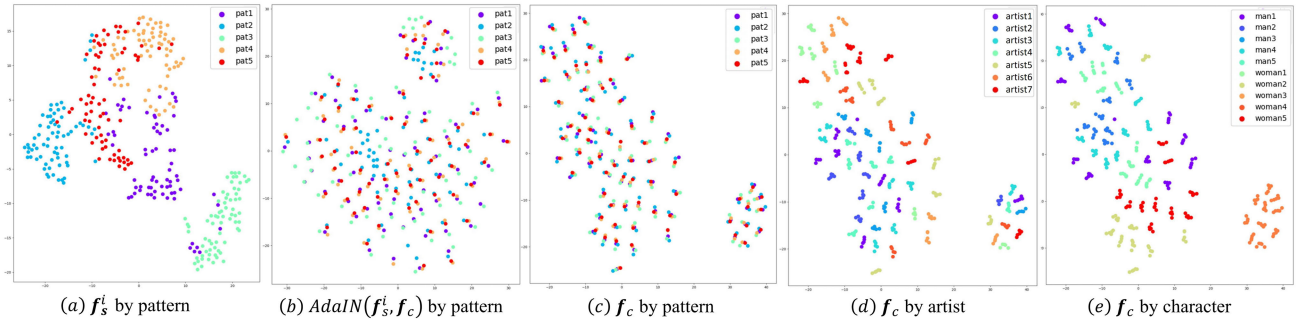


FIGURE 6. Latent feature visualization results. (a) pattern-dependent style features f_s^i are clustered by pattern i. (b) combined feature $AdaIN(f_s^i, f_c)$ is dispersed compared to f_c. (c) f_c cannot be clustered by pattern i, and each group covers all patterns. (d,e) f_c are labeled by the artists/characters. The latent code tends to concentrate according to the artist but is smoothly distributed in terms of the characters.

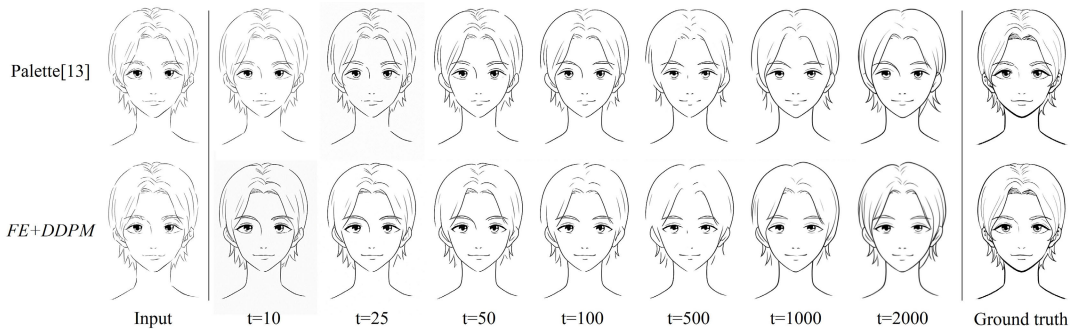


FIGURE 7. Image generation results when prediction step number t increases for pat.2.

of stroke style features can be separated successfully from the content features.

Furthermore, we analyze the dispersion of latent code clusters by analyzing the t-SNE latent maps shown in Fig. 5. In each map, $\boldsymbol{f}_c$ are clustered on the mean while combined features $AdaIN(\boldsymbol{f}_s^i, \boldsymbol{f}_c)$ are clustered around them. This result shows that the dispersion of the latent codes increases when patterns of stroke style are added to the content. This shows that once the pattern of style is added to the content features, our method makes the combined features $AdaIN(\boldsymbol{f}_s^i, \boldsymbol{f}_c)$ separate from each other.

B. IMAGE GENERATION RESULTS

We evaluate the quality of the synthesized images yielded by the FE step and the $FE + DDPM$ step both quantitatively and qualitatively. First, we choose some metrics to evaluate the results in terms of absolute accuracy and image quality. Second, we sample some results from the $FE + DDPM$ step and choose the final results by experiments with a discussion on sampling speed and quality tradeoffs. Third, generation results of the FE step and the $FE + DDPM$ step are compared to state-of-the-art works. Finally, we discuss the different refinement performances yielded by different style patterns.

1) EVALUATION METRICS

We evaluated the L2 distance, Learned Perceptual Image Patch Similarity (LPIPS), and Structural Similarity Index (SSIM) score between the generated image and the ground truth, see Table 4. These evaluation criteria judged our experimental results from different perspectives.

L2 distance measures the pixel or feature differences between two images. Learned Perceptual Image Patch Similarity (LPIPS) quantifies the perceptual similarity between images based on higher-level features and aligns with human perception. These metrics are useful for evaluating image similarity, assessing the quality of generated images, and understanding the perceptual differences between images. In addition, the Structural Similarity Index (SSIM) is a metric that evaluates image similarity by considering both structural and perceptual aspects. It measures how well the structure, contrast, and luminance are preserved between two images; it provides a comprehensive measure of similarity and quality.

2) DISCUSSION ON SAMPLING SPEED AND QUALITY TRADEOFFS

Discussion on previous work [27] draws to a conclusion that the best prediction steps are always the same as the training noise level T, which always costs a long calculation time. When applying the diffusion model to the issue of light refinement in line arts, we discuss the length of the prediction step t. As sampling step number t increases, the L2 loss results of our $FE + DDPM$ step results become worse, while those of diffusion methods [13], become better, see Table. 3. The best result of our work is when t equals 25, while DDPM's results need a larger t (around 50 to 100). The image generation results also show the same phenomenon, shown in Fig. 7. Smaller t adds less noise to the original images which

TABLE 3. Generation performance in terms of prediction step number t.

	t=25	t=50	t=100	t=500	t=2000
L2↓					
Palette [13]	35.874	35.704	**35.427**	35.513	39.123
Ours: $FE + DDPM$	**32.078**	32.550	32.901	33.928	38.956
LPIPS↓					
Palette [13]	0.075	**0.074**	0.077	0.094	0.091
Ours: $FE + DDPM$	**0.070**	0.072	0.077	0.102	0.090
SSIM↑					
Palette [13]	0.844	**0.845**	0.844	0.835	0.817
Ours: $FE + DDPM$	**0.846**	0.845	0.842	0.830	0.811

ensures that the line changes are not so large as to prevent reconstruction of the stroke style features, merely to refine them slightly. Furthermore, DDPM needs a larger t indicating that the *FE* step achieves good refinement performance on lines with different patterns of stroke styles, which can offset DDPM's initial prediction steps. The disentanglement extracts common features among different patterns, which helps the *DDPM* step easily refine different lines regardless of the style patterns.

The experiment results show that adding the *FE* step to the *DDPM* step helps to generate high-quality results with smaller prediction steps compared to pure DDPM's requirement for more steps. It can be inferred that the *FE* step can replace the early-stage computation in the low-dimensional space of the diffusion model, thereby reducing the sampling time. This means that our approach achieves efficient and rapid sample generation without sacrificing quality.

3) STATE-OF-THE-ART COMPARISONS

Line art refinement techniques need the technique points on three topics: style, content, and suitability to treat the line arts. Since there is no research that covers all these topics, We compared our work with six previous standard works in recent years that contain one or two topics. Sketch simplification techniques [6], [7] can only change the shape of contents, and they are specific to line art. Style transfer method [2] focuses on changing the style. Image-to-image [3] and Multi-domain disentanglement [10] can handle different styles and content, but it is not specifically designed for line art. Furthermore, the current image generation method using DDPM models [13] is the best one that is good at treating content in detail. We choose the best results of $t = 25$ as our $FE + DDPM$ final results and $t = 50$ as the best results for the original DDPM method [13]. As shown in Table 4, the best results are marked and indicate that our results of *FE* step and $FE + DDPM$ step are superior to those of previous works [2], [3], [6], [7], [10], [13] across all the evaluation metrics we have examined.

In addition, the visual comparison shown in Fig. 8 shows that $FE + DDPM$ step outperforms the previous works. We can see that the synthesized results from *FE* step are blurred while those from $FE + DDPM$ step are sharp. Our method first removes the stroke style that yields blurred images, and then, based on the blur results, the $FE + DDPM$ step creates corrected clear line artwork. The rows follow

TABLE 4. Generation performance on test data. Bold and underline indicate the best and the second best result.

Method	L2↓	LPIPS↓	SSIM↑
AdaIN [2]	36.866	0.078	0.820
CycleGAN [3]	38.117	0.080	0.839
DMIT [10]	38.034	<u>0.077</u>	0.843
Mastering Sketching [6]	38.034	<u>0.077</u>	0.843
Smart Inker [7]	45.621	0.101	0.826
Palette [13](t=50)	35.704	0.074	<u>0.845</u>
Ours: FE	**28.641**	0.129	<u>0.845</u>
Ours: $FE + DDPM$ (t=25)	<u>32.037</u>	**0.070**	**0.846**

TABLE 5. User study: 20 volunteers evaluate the generation results.

	closest to GT
AdaIN [2]	8.0 %
CycleGAN [3]	11.5 %
DMIT [10]	11.0 %
Mastering Sketching [6]	7.0 %
Smart Inker [7]	1.0 %
Palette [13](t=50)	<u>23.5 %</u>
Ours: $FE + DDPM$ (t=25)	**38.0 %**

the order of pat.1 to pat.5. Analyzing the results from the perspective of uniformity, the key aspect of the final results is the ability of our method to generate images in a unified stroke style from different patterns of input images. From the comparison of the generated images across different patterns, it is evident that only $FE + DDPM$ can successfully refine the five different style patterns. Furthermore, $FE + DDPM$ results are smoother, more coherent, and most similar to the ground truth.

We also add a user study on generation quality evaluation by evaluating the generated images. The sample generated covers all artists and patterns from the validation data. 20 volunteers evaluate the sampled generated results 10 times to evaluate the closest generated image to the ground truth, shown in Table.5. Volunteers chose the results produced by our proposed method as the ones that aligned most closely with the ground truth, which earned the same results as other quantitative metrics. All these results demonstrate that our proposed networks can generate images that present the stroke styles of professional artists.

4) DISCUSSION ON QUALITY BY PATTERNS

We also compare the refinement performances yielded by different patterns. As shown in Fig. 8, compared to the previous methods (2nd-7th columns), our method is able to change the shape of the curves or the connection of the lines. Our method smooths the redundant short lines better than the previous method for pat.1 *fuzzy*. Moreover, broken lines pat.2 *casual* can be reconnected, especially the outline of the chin and the neck, unlike the previous methods. In addition, our method attenuates more lines for pat.3 *overlap*. Our method can create lines with much more strength variance pat.4 *no strength*. Furthermore, our method modifies sharp curves pat.5 *hand-shaking* better than the previous methods. These results show that our network reshapes images toward the ground truth style regardless of the detailed strokes in the input line image. Once the *FE*

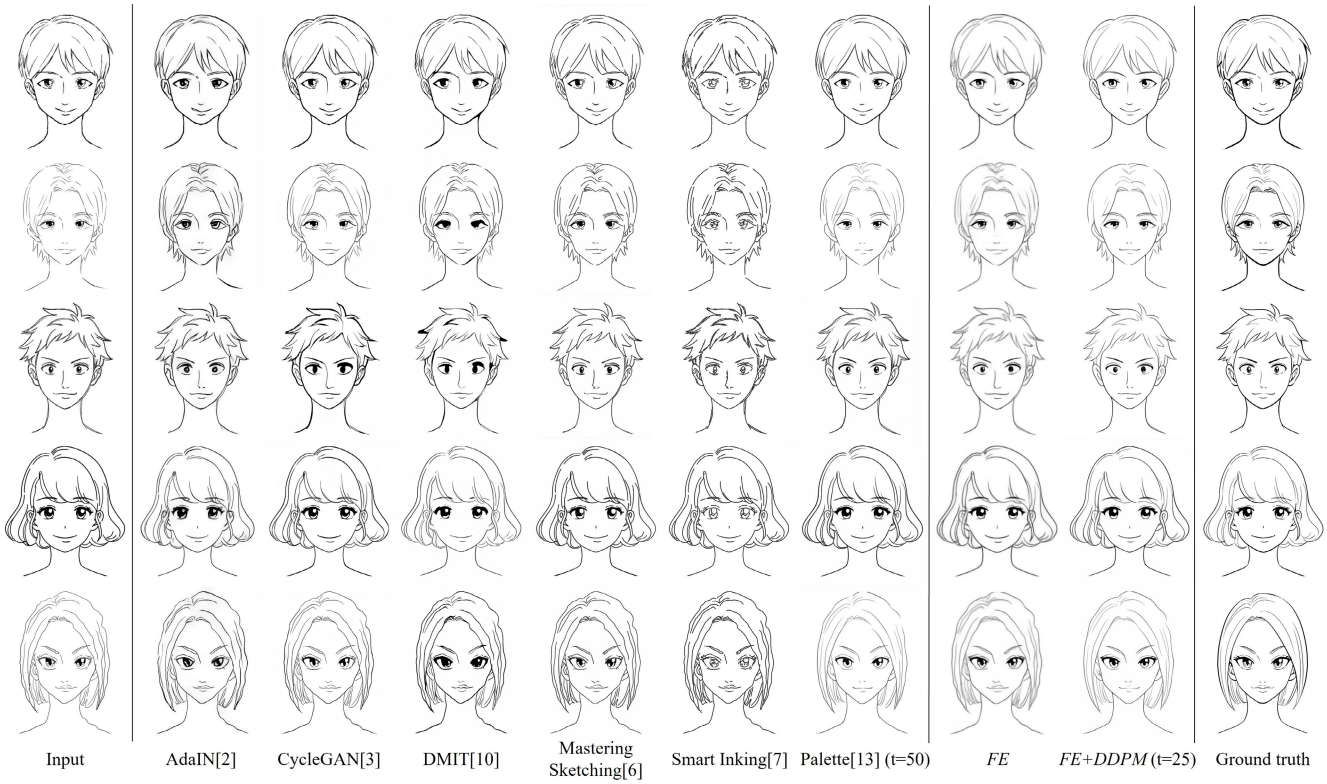


FIGURE 8. **Image generation results for test data. Our method smoothed the redundant short lines better than previous methods for pat.1 *fuzzy*. Broken lines pat.2 *casual* are reconnected, especially the outline of the chin and the neck. Our method attenuated more lines given pat.3 *overlap*. Our method is able to create lines with much more strength variance pat.4 *no strength*. Furthermore, our method modified the severe curves of pat.5 *hand-shaking* better than the previous work.**

TABLE 6. **Image generation results by patterns.**

Pattern	Method	L2↓	LPIPS↓	SSIM↑
pat.1	Palette [13](t=50)	30.199	0.062	**0.875**
	Ours: $FE + DDPM$(t=25)	**28.549**	0.062	0.871
pat.2	Palette [13](t=50)	36.864	0.094	0.818
	Ours: $FE + DDPM$(t=25)	**33.958**	**0.086**	**0.820**
pat.3	Palette [13](t=50)	38.282	0.078	0.830
	Ours: $FE + DDPM$(t=25)	**32.974**	**0.073**	**0.835**
pat.4	Palette [13](t=50)	37.034	0.068	**0.862**
	Ours: $FE + DDPM$(t=25)	**32.152**	**0.061**	0.861
pat.5	Palette [13](t=50)	36.104	0.070	0.840
	Ours: $FE + DDPM$(t=25)	**32.775**	**0.068**	**0.841**

step is added to DDPM [11], the $FE + DDPM$ results more closely approach the ground truth, especially for pat.2 and pat.4. Table. 6 reveals the superiority of our proposed method over Palette [13] regardless of pattern. These patterns exhibit greater dependency of stroke position on the content, which is hard to offset. These results show that our method can better handle

Overall, based on the above qualitative and quantitative results, our proposed network that disentangles the stroke style and content enhances the line refinement performance, highlighting its strengths in capturing accurate shapes and preserving structural similarity across a variety of patterns. Moreover, our methods add more informative details resulting in the results being more similar to professional artwork.

VI. CONCLUSION

We have created an original high-quality paired line artwork dataset that covers the six patterns of stroke styles. Our proposed approach can disentangle the pattern-dependent stroke style features which helps the novice by providing expert-level hints. At the same time, extracted pattern-independent content features achieve correct line art refinements. Moreover, the proposal's *DDPM* step increases the resolution and quality of generated images.

This paper offers the following key contributions: Creation of a novel cataloged line art dataset that encompasses various stroke styles; Development of specialized models with enhanced disentanglement capabilities to effectively separate distinct style features and precise content features from line drawings, which enables style-specific recommendations; Generation of high-quality refined line art with intricate details, resulting in visually appealing output.

Experiments on the proposal's classification accuracy, disentanglement performance, and generation quality demonstrate the efficiency of pattern-dependent feature removal and the superiority of our method over previous methods. Regardless of the stroke style patterns of the input image, our method successfully suppresses its characteristics and modifies them to more closely approach the target stroke style. Moreover, disentanglement of the common features can

replace the early stage process of the diffusion model, which results in reducing the sampling time.

To make the proposal suitable for more arbitrary artworks, we plan to develop adaptive methods that can represent a wider variety of line styles.

REFERENCES

[1] M. Morinaga, *An Introduction to Drawing for Hetappi: What does 'drawing a line' mean!?* Japan: Impress Corporation, 2018.

[2] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[4] X. Liu, W. Wu, H. Wu, and Z. Wen, "Deep style transfer for line drawings," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 353–361.

[5] E. Simo-Serra, S. Iizuka, K. Sasaki, and H. Ishikawa, "Learning to simplify: Fully convolutional networks for rough sketch cleanup," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, Jul. 2016.

[6] E. Simo-Serra, S. Iizuka, and H. Ishikawa, "Mastering sketching: Adversarial augmentation for structured prediction," *ACM Trans. Graph.*, vol. 37, no. 1, pp. 1–13, Feb. 2018.

[7] E. Simo-Serra, S. Iizuka, and H. Ishikawa, "Real-time data-driven interactive rough sketch inking," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–14, Jul. 2018.

[8] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 2, pp. 5–9, 2016.

[9] X. Peng, Z. Huang, X. Sun, and K. Saenko, "Domain agnostic learning with disentangled representations," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)* (Proceedings of Machine Learning Research), vol. 97, Jun. 2019, pp. 5102–5112.

[10] X. Yu, Y. Chen, S. Liu, T. Li, and G. Li, "Multi-mapping image-to-image translation via learning disentanglement," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020, *arXiv:2006.11239*.

[12] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2023.

[13] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *Proc. Special Interest Group Comput. Graph. Interact. Techn. Conf.*, Aug. 2022, pp. 1–10.

[14] T. Igarashi, S. Matsuoka, S. Kawachiya, and H. Tanaka, "Interactive beautification: A technique for rapid geometric design," in *Proc. 10th Annu. ACM Symp. User Interface Softw. Technol.*, 1997, pp. 105–114.

[15] X. Liu, T.-T. Wong, and P.-A. Heng, "Closure-aware sketch simplification," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–10, Nov. 2015.

[16] Y. Liu, X. Li, P. Bo, and X. Gao, "Sketch simplification guided by complex agglomeration," *Sci. China Inf. Sci.*, vol. 62, no. 5, p. 52105, May 2019.

[17] F. Cole, D. DeCarlo, A. Finkelstein, K. Kin, K. Morley, and A. Santella, "Directing gaze in 3D models with stylized focus," in *Proc. Eurograph. Symp. Rendering*, Jun. 2006, pp. 377–387.

[18] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015, *arXiv:1508.06576*.

[19] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," 2016, *arXiv:1603.08155*.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirzaand, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014.

[21] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," 2016, *arXiv:1609.04802*.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[23] F. Xie, M. Goto, and H. Seshimo, "Disentangling defect-aware aesthetic features for line art modification," *Special Interest Group Comic Comput.*, 2022. [Online]. Available: https://drive.google.com/file/d/1WxVOYb_6GluyFva4moWVNiR3GP0UPX6L/view

[24] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "ILVR: Conditioning method for denoising diffusion probabilistic models," 2021, *arXiv:2108.02938*.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

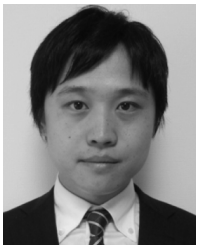
[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.

[27] K. Pandey, A. Mukherjee, P. Rai, and A. Kumar, "DiffuseVAE: Efficient, controllable and high-fidelity generation from low-dimensional latents," 2022, *arXiv:2201.00308*.

[28] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

FANGLU XIE received the B.E. degree in automation from Southeast University, Nanjing, China, in 2016, and the M.E. degree in information, production and systems from Waseda University, Fukuoka, Japan, in 2017.

In 2018, she joined Nippon Telegraph and Telephone Corporation (NTT). She is currently a Researcher with NTT Human Informatics Laboratories. Her research interests include image processing, computer vision, machine learning, and generative artificial intelligence.



MOTOHIRO TAKAGI (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Keio University, in 2009, 2011, and 2020, respectively.

In 2011, he joined Nippon Telegraph and Telephone Corporation (NTT). He is currently a Researcher with NTT Human Informatics Laboratories. His research interests include human behavior understanding through machine learning, computer vision, and natural language processing.

HITOSHI SESHIMO received the B.E. and M.E. degrees in mechanical engineering from Waseda University, Tokyo, Japan, in 1995 and 1997, respectively.

In 1997, he joined Nippon Telegraph and Telephone Corporation (NTT). His research interests include computer-aided instruction, web-based learning, content distribution and navigation systems, geographical information services, and cybernetics.

YUSHI AONO received the B.E., M.E., and Ph.D. degrees from Osaka University, Osaka, Japan, in 1994, 1996, and 1999, respectively.

Since 1999, he has been with Nippon Telegraph and Telephone Corporation (NTT), Tokyo, Japan, engaged in research on speech recognition and speech synthesis. He is currently the Manager with the Cybernetics Laboratory, NTT Human Informatics Laboratories. He is a member of the Information Processing Society of Japan and the Institute of Electronics, Information and Communication Engineers.