SKETCH2MANGA: SHADED MANGA SCREENING FROM SKETCH WITH DIFFUSION MODELS

Jian Lin[1], *Xueting Liu*[*1], *Chengze Li*[1], *Minshan Xie*[2], *Tien-Tsin Wong*[3]

[1]School of Computing and Information Sciences, Saint Francis University
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[3]Department of Data Science and Artificial Intelligence, Monash University

ABSTRACT

While manga is a popular entertainment form, creating manga is tedious, especially adding screentones to the created sketch, namely *manga screening*. Unfortunately, there is no existing method that tailors for automatic manga screening, probably due to the difficulty in generating shaded high-frequency screentones of high-quality. Classic manga screening approaches generally require user input to provide screentone exemplars or a reference manga image. Recent deep learning models enable automatic generation by learning from a large-scale dataset. However, the state-of-the-art models still fail to generate high-quality shaded screentones due to the lack of a tailored model and high-quality manga training data. In this paper, we propose a novel sketch-to-manga framework that first generates a color illustration from the sketch and then generates a screentoned manga based on the intensity guidance. Our method significantly outperforms existing methods in generating high-quality manga with shaded high-frequency screentones.

Index Terms— manga generation, manga screening, sketch-to-manga

1. INTRODUCTION

Manga is a popular entertainment form that presents stories through black-and-white illustrations. Due to the lack of color, screentones (black-and-white patterns) are used to present shading and textures. To produce manga pages, artists usually first create a sketch to depict the contour of the objects and then apply screentones to the sketch at proper places with proper shading. While drawing the contour of the objects comes from the creative thoughts of the artists, the screening process is tedious and repetitive, with the goal of achieving natural yet visually striking imitations of color shades and conveying semantic meanings. An automatic manga screening approach is highly desirable to reduce labor cost, yet unfortunately, this task is extremely challenging because of the large information gap between the sketch domain and the manga domain. The selection and application of screentones pose significant challenges due to the bitonal nature of manga, requiring careful aesthetic and semantic considerations. Additionally, it is not easy to generate black-and-white screentones of high-frequency regular patterns, even with recent deep learning models.

In this work, we propose a novel diffusion-based framework for automatic manga screening. The key of our method is to establish color illustration as an intermediary between sketch and manga. Utilizing color illustration as a proxy simplifies the complex screening task by breaking it down into two more manageable subtasks. In the first step, we generate a color illustration from the sketch that carries shading information with a text-to-image diffusion model with line conditioning. In the second step, we generate a manga image with high-frequency screentones from the shaded color illustration. Considering the ultimate goal of mimicking color shading with screentones, the transition from color illustrations to manga is both intuitive and meaningful. To our knowledge, there is currently no fully automatic approach designed specifically for manga screening from sketch. Existing methods primarily concentrate on photo-to-manga conversion techniques [1, 2, 3, 4, 5], which still present significant domain gaps, or on reference-based screening methods [6, 7] that require additional reference input. While recent advancements in diffusion-based image synthesis have demonstrated impressive capabilities in the quality and controllability of generation through text [8] or other conditional inputs [9], diffusion models inherently struggle to produce high-quality screentones. This limitation is mainly attributed to the lack of high-fidelity manga data available for training latent diffusion models, including both the variational autoencoder (VAE) and the U-Net model, resulting in a propensity to generate visuals with low-frequency characteristics.

In light of these challenges during manga generation, our framework first improves the stable diffusion model by finetuning its components with a newly curated high-quality manga dataset. Specifically, we revisit and reformulate the training objectives of its VAE model to better adapt to the

*Corresponding author (e-mail: tliu@sfu.edu.hk). This work was supported by the Research Grants Council of the Hong Kong SAR, China (Project No. UGC/FDS11/E01/21).

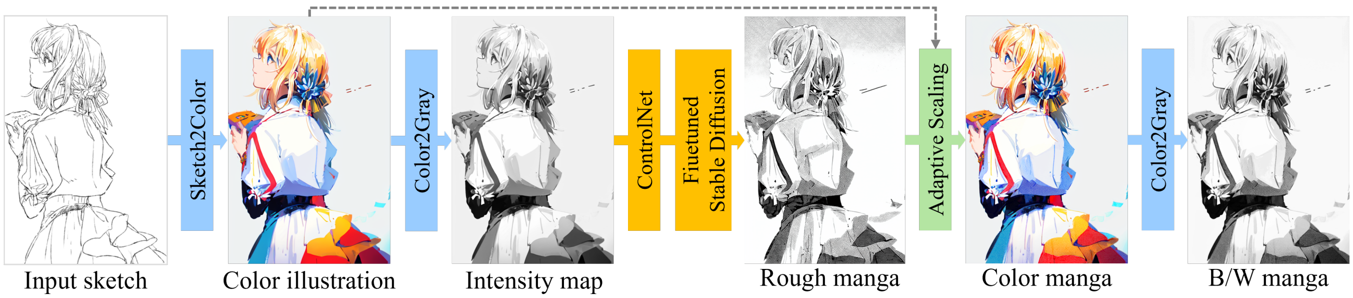


Fig. 1. System overview. Given an input sketch, our system first generates a color illustration from the sketch and then generates a rough manga image using a diffusion model conditioned on the intensity map of the color illustration. An adaptive scaling process is further applied to generate the final manga image.

intricacies of high-frequency screentones. For the denoise diffusion U-Net model, we utilize the intensity of color illustration as the additional controlling factor besides text prompt to ensure that the screentones generated for the manga align with the shading of the color illustration. Despite the advancements achieved through finetuning, the manga images occasionally exhibit minor structural distortion and unnatural shading, a byproduct of the denoise diffusion process. To rectify these issues, we develop an adaptive scaling method that effectively incorporates the generated screentones into the color illustrations, culminating in a manga output that is both structurally sound and aesthetically pleasing.

We validate the effectiveness of our method by comparing with both sketch-to-manga methods and illustration-to-manga methods. The results show that our method significantly outperforms existing methods. Our contributions are summarized as follows:

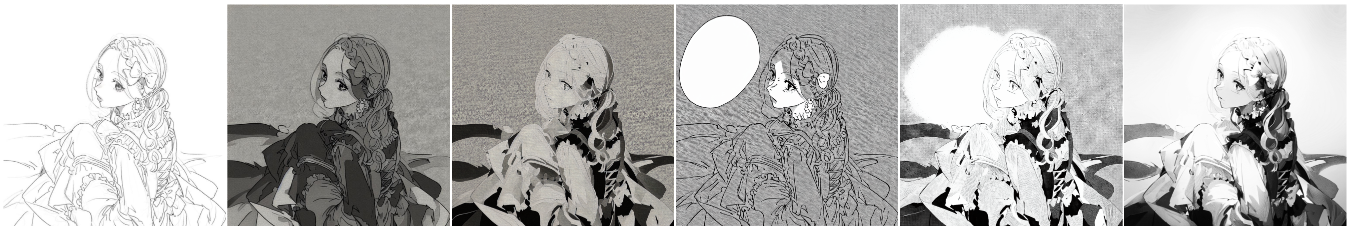
- We propose a novel sketch-to-manga framework to generate high-quality manga with shaded high-frequency screentones by using color illustration as an intermediary to bridge the gap between line drawing and screened manga.
- We finetune a diffusion model with a tailored objective and intensity conditioning to generate high-quality shaded high-frequency screentones.
- We propose an adaptive scaling method to integrate screentones into a color illustration to obtain screentoned manga.

2. RELATED WORK

Sketch-to-manga While various sketch colorization methods (e.g., [10, 11, 12]) have been proposed, very few attempts have been made in adding screentones to sketches. To enable automatic manga screening, [13] proposed to first predict a screentone category label for each pixel via a neural network and then apply the corresponding screentone to each screentone category area. Due to the unstable performance of the classification network, the generated screentone category label map is error-prone and the generated screentones are unnatural and unshaded. Some reference-based manga screening methods have been proposed to transfer the screentones from a manga image to a line drawing [6, 7], but these methods rely on additional information to apply the screentones and synthesize the manga image, which are by nature not automatic. In this paper, we propose a fully automatic method without relying on any additional information.

Illustration-to-manga Compared with the limited research efforts in sketch-to-manga, significantly more work has been done in converting color illustration to manga. The first photo-to-manga approach [1] proposed to analyze the correspondence between color and screentone, and then convert color to screentone based on manually selected screentone exemplars. To avoid manual screentone selection, some applications [2, 3] have been made to automate screentone generation; yet without screentone exemplars, their screentones generated are usually unnatural and of low quality due to lack of semantic consistency with the original. With the development of deep learning, ScreenVAE [4], has been proposed to allow the inter-transfer of color illustration and manga via variational inference. However, this screentone encoder still fails to generate high-quality screentones, especially in color-rich regions. Later, another deep learning approach [5] has been proposed to estimate the screentone filling strategy via segmentation and then generate high-frequency screentones based on the segmentation map and screentone exemplars. While this method achieves state-of-the-art performance, the generated manga still cannot preserve the shading of the original image and exhibits overall low-contrast. In comparison, we propose a method that can generate high-quality manga with well-shaded high-frequency screentones.

Diffusion Models and ControlNet Image generation has garnered significant attention in deep learning due to the widespread demand to synthesize high-quality photography. Among existing models, diffusion models have proven successful in image generation. These models are trained to iteratively denoise noise-corrupted samples drawn from a Gaussian distribution, resulting in high-quality generation. The denoising process is performed mainly in latent space [8] for computational efficiency, and the diffusion modules are often incorporated with cross-attention modules that process



(a) Input sketch (b) Original cond. on sketch (c) Original cond. on intensity (d) Finetuned cond. on sketch (e) Finetuned cond. on intensity (g) Output manga based on (e)

Fig. 2. Ablation on latent diffusion. The original diffusion model fails to generate high-frequency screentones without finetuning. After finetuning, conditioning on sketch still leads to flat shading, while conditioning on intensity generates shaded high-frequency screentones.

text embeddings to allow text-to-image generation. The controlNet [9] conditions the latent diffusion model on various conditioning signals, such as sketch, to ensure that the generation process closely follows the conditioning signal. However, the latest off-the-shelf VAE model for latent diffusion still generates distorted and blurry screentones when used to create manga images from the latent space due to significant information loss [14]. Furthermore, directly generating a manga image using the stable diffusion model conditioned on sketch leads to bland results due to the lack of information in the sketch guidance. In this paper, we propose to finetune the stable diffusion model with high-quality manga data. We also propose a two-step approach to first estimate the color illustration of the sketch and then generate a high-quality manga image with rich screentones via the finetuned stable diffusion model conditioned on the intensity map.

3. METHODOLOGY

Fig. 1 presents the system architecture of our framework.

3.1. Sketch to color illustration

We first propose to generate a color illustration from the input sketch with standard diffusion models fine-tuned on color illustrations from the community, denoted as *Sketch2Color* in Fig. 1. Specifically, we use MeinaPastel[1], a diffusion model finetuned on colored illustration, as our colorization model, and control_v11p_sd15_lineart[2] to condition the denoising process on line-art sketches. We choose the same set of text prompts[3] for all generations. The generated color illustration is richer in information than the original sketch, encompassing elements such as color and shading. Subsequently, we obtain the intensity values derived from the grayscale version of the color illustration to guide the manga generation with another diffusion model, denoted as *Color2Gray* in Fig. 1.

[1]https://civitai.com/models/11866/meinapastel

[2]https://github.com/lllyasviel/ControlNet-v1-1-nightly

[3]Positive: *masterpiece, best quality, solo, illustration, simple background*; Negative: *nsfw, nude, lowres, bad anatomy, bad hands, worst quality, low quality, normal quality, jpeg, jpeg artifacts*.

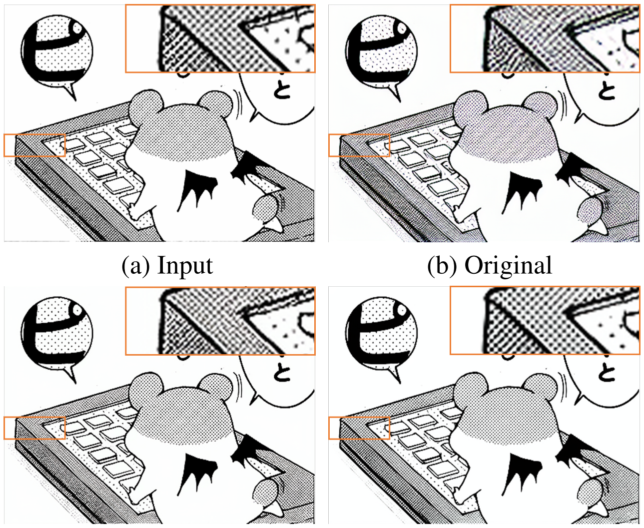
3.2. Intensity-guided manga generation

To create manga images with screentones from a sketch, we can use a pre-trained ControlNet [9] with line conditioning to help guide the generation process with a stable diffusion model [8]. However, there is a challenge – many manga images found online are of low quality, often due to being scanned or being compressed for sharing. This means that high-quality screentone images are rare during model pretraining, which leads to low-quality generated manga images. Moreover, we discover that conditioning ControlNet on the grayscale intensity map derived from the colored illustration obtained in the previous step results in a broader range of screentone variations that align more closely with the concept of shading, as opposed to conditioning solely on the sketch.

To obtain a large number of high-resolution manga training samples, we scrape 186k high-resolution manga images from the Internet and further restore the high-resolution versions of the Manga109 dataset [15] using a manga restoration method [16] to obtain 103k more high-quality manga images. Altogether 289k high-resolution manga images are collected. Since the prepared manga images are usually full pages containing multiple manga panels, the panels are cropped using Grounded DINO (with query *panel*) as training data. The prompts used for text-to-image generation are generated by a SwinV2 tagger[4] pretrained on Danbooru [17] dataset.

With the prepared training data, we finetune both denoise diffusion U-Net and the VAE decoder components in stable diffusion. The effectiveness of finetuning the latent diffusion is shown in Fig. 2. The original latent diffusion model fails to generate high-frequency screentones due to the lack of training data, no matter the generation is conditioned on a sketch or an intensity map (Fig. 2(b)&(c)). Finetuning the latent diffusion enables the generation of high-frequency screentones, but conditioning on sketch still leads to bland unshaded screentones (Fig. 2(d)). With intensity conditioning, our finetuned latent diffusion generates manga with shaded high-frequency screentones (Fig. 2(e)). However, finetuning the latent diffusion alone still fails to generate visually pleas-

[4]https://github.com/SmilingWolf/SW-CV-ModelZoo



(a) Input (b) Original

(c) Finetuned with LPIP loss (d) Finetuned without LPIP loss

Fig. 3. Ablation on VAE decoder. The original VAE decoder fails to generate visually pleasant screetones. Finetuning the VAE decoder without LPIP loss produces the best results.

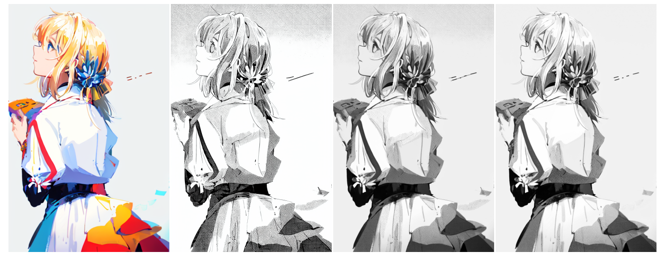
ant screentones, as shown in Fig. 3(b). So we further finetune the decoder by following the paradigm of sd-vae-ft-mse[5] that uses a weighted sum of adversarial loss [18], MSE loss, and LPIP loss [19] as the objective function. Unfortunately, the result is still unsatisfactory after the finetuning (Fig. 3(c)). We find that LPIP loss degrades screentone reconstruction, probably due to the domain shift between manga and natural images. Therefore, we further propose to remove the LPIP loss from the objective function, which significantly enhances the quality of the generated screentones (Fig. 3(d)).

3.3. Adaptive scaling

Our finetuned intensity-conditioned diffusion model produces manga with shaded high-frequency screentones, but also with subtle structure distortion and unnatural shading introduced by the diffusion process (as shown in the area around the eyes in Fig. 4(b)), we propose a color-screentone integration scheme to integrate the generated color illustrations and the generated rough manga with shaded high-frequency screentones to produce the final manga image.

Given an input sketch I_s, we denote the generated color illustration as I_c and its grayscale version as I_g. The rough manga image with shaded high-frequency screentones generated by the diffusion model is denoted as I_r. We first apply K-means clustering to the generated color illustration I_c to obtain a set of regions where pixels in each region are connected and with similar colors. For each identified region R, the distribution of screentones in the generated rough manga

[5]https://huggingface.co/stabilityai/sd-vae-ft-mse



(a) Color illustration (b) Rough manga (c) Scaling on L (d) Scaling on S

Fig. 4. Example of adaptive scaling on lightness (L) and saturation (S) respectively. Scaling on saturation achieves better results by preserving the shading of the color illustration.

image is utilized to decide how likely this region should be screentoned. We propose that if the standard deviation of the region in the generated rough manga image σ^R is large, it is very likely that this region R should be screentoned, and vice versa. Based on this assumption, we calculate a scaling range $[s^R_{low}, s^R_{high}]$ defined by two scaling factors s^R_{low} and s^R_{high} for each region R based on its standard deviation σ^R as

$$\begin{aligned} s^R_{low} &= 1 - w_{low} \cdot \sigma^R \\ s^R_{high} &= 1 + w_{high} \cdot \sigma^R \end{aligned}$$

Here, the scaling factors s^R_{low} and s^R_{high} indicate the lowest bound and the highest bound of the scaling range respectively. w_{low} and w_{high} are two weighting parameters and empirically set to 0.08 and 0.16 in all our experiments. With this pair of region-based scaling factors s^R_{low} and s^R_{high}, we are able to scale the colors of all pixels inside the region R in the color illustration I_c in an adaptive way.

We propose two adaptive scaling strategies by scaling the colors on the lightness channel L_c and saturation channel S_c respectively. In the following, adaptive scaling on the saturation channel is described, while adaptive scaling on the lightness channel can be obtained in a similar manner. For each region R of I_r, we denote the minimum and maximum intensity values of this region as $\min(I^R_r)$ and $\max(I^R_r)$ respectively. The value of a pixel p in the saturation channel S_c is updated by multiplying a scaling factor s, i.e., $S_c[p] = S_c[p] \cdot s$, where s is calculated as

$$s = s^R_{high} - \frac{I_r[p] - \min(I^R_r)}{\max(I^R_r) - \min(I^R_r)} \cdot \left(s^R_{high} - s^R_{low}\right)$$

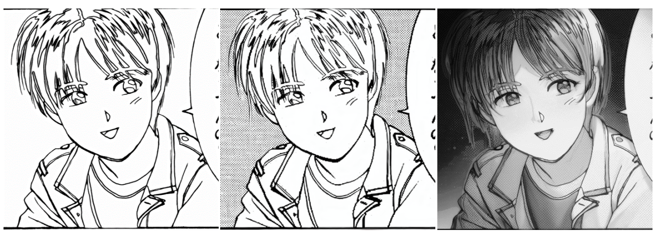
Fig. 4 shows an example in which both the results of the scaling on the lightness channel and the scaling on the saturation channel are presented. We can observe that scaling on the lightness channel can better preserve the generated screentones, while scaling on saturation can better preserve the shading in the color illustration. It is observable that the latter one provides more natural shading, similar to that in real-world manga. For example, a flat low-contrast region is

Input pix2pix [20] Screentone Synthesis [13] Ours

Fig. 5. Comparisons with state-of-the-art sketch-to-manga methods.

Sketch-to-manga	Preferred	Illustration-to-manga	Rank 1	Rank 2	Rank 3
Screentone Synthesis [13]	16.2%	ScreenVAE [21]	3.0%	6.4%	90.6%
Ours	**83.8%**	Mimicking Manga [22]	32.9%	66.5%	0.6%
		Ours	**64.1%**	27.1%	8.8%

Table 1. User studies on sketch-to-manga and color-to-manga.



(a) Input (b) Without intermediary (c) With intermediary

Fig. 6. Ablation on importance of using color illustration as intermediary between line drawing and screened manga.

less likely to be screened, and the tone intensity and variations of the screentones in a region should align with the shading in this region. So, region-based adaptive scaling on the saturation channel is adopted in obtaining all our examples.

4. RESULTS

To validate the effectiveness of our method, we compare with the state-of-the-art sketch-to-manga methods and illustration-to-manga methods. For sketch-to-manga, only one automatic method has been proposed so far, namely *Screentone Synthesis* [13]. We also compare with the image-to-image translation model pix2pix [20]. For illustration-to-manga, we compare with two state-of-the-art methods, including ScreenVAE [21] and Mimicking Manga [22]. We did not include quantitative evaluations, since there are no established metrics suitable for this purpose. Metrics like LPIP and FID, which rely on DNNs trained on labeled natural images, are not applicable to manga due to the absence of comparable datasets.

Visual Comparisons Fig. 5 shows the comparisons with existing sketch-to-manga methods. Pix2pix [20] generates blurred unshaded screentones with unnatural placement. Screentone Synthesis [13] generates relatively high-frequency screentones, but are still unshaded with unnatural placement. In comparisons, our method generates high-quality manga images with high-quality shaded high-frequency screentones.

Fig. 7 shows the comparisons with existing color-to-manga methods. ScreenVAE [21] generates high-frequency screentones with proper shading transitions in flat or almost flat color regions. However, this model faces challenge in regions with rich colors, leading to frequently occurring black regions in the generated manga image. Mimicking manga [22] generates high-quality screentones since they rely on screentone exemplars to synthesize the screentones in the output image. However, the screentones are unnaturally shaded, leading to a consistent low contrast manga output. In contrast, our method achieves significantly better results than existing methods, even for color illustration inputs.

We further show an example in Fig. 6 to demonstrate the importance of the intermediate coloring stage. Without the coloring stage, the generated screentones are limited and unshaded. In sharp comparison, using color illustration as an intermediary leads to high-quality shaded screentones.

User Studies We conduct user studies with 20 unseen sketches for sketch-to-manga and 20 unseen illustrations for illustration-to-manga from Danbooru [17]. 17 participants are invited. For sketch-to-manga, since there is only one competitor, the users are asked to choose their preferred one from the two results. For illustration-to-manga, since there are two competitors, the users are asked to rank the three results based on their preferences. Table 1 shows the results where our method significantly outperforms the rest.

(a) Input (c) ScreenVAE [21] (b) Mimicking Manga [22] (d) Ours

Fig. 7. Comparisons with state-of-the-art illustration-to-manga methods.

5. CONCLUSION

In this paper, we proposed a novel sketch for the manga framework that utilized color illustration as the intermediary. A tailored manga generation diffusion model conditioned on intensity is proposed by finetuning on high-resolution manga data with a tailored loss function. An adaptive scaling method is further proposed to integrate the generated rough manga with high-quality screentones into the color illustration.

The proposed method could be further extended to add screentones to a full manga page or a line cartoon. We would further explore such an application while maintaining the consistency among different panels in a manga page or among consecutive frames in a line cartoon animation.

6. REFERENCES

[1] Yingge Qu, Wai-Man Pang, Tien-Tsin Wong, and Pheng-Ann Heng, "Richness-preserving manga screening," *ACM Transactions on Graphics (TOG)*, vol. 27, no. 5, pp. 1–8, 2008.

[2] Hua Li and David Mould, "Content-sensitive screening in black and white," in *International Conference on Computer Graphics Theory and Applications*, 2011, vol. 2, pp. 166–172.

[3] Zhipeng Wu and Kiyoharu Aizawa, "Mangawall: Generating manga pages for real-time applications," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 679–683.

[4] Minshan Xie, Chengze Li, Xueting Liu, and Tien-Tsin Wong, "Manga filling style conversion with screentone variational autoencoder," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–15, 2020.

[5] Lvmin Zhang, Xinrui Wang, Qingnan Fan, Yi Ji, and Chunping Liu, "Generating manga from illustrations via mimicking manga creation workflow," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5642–5651.

[6] Huisi Wu, Ziheng Ma, Wenliang Wu, Xueting Liu, Chengze Li, and Zhenkun Wen, "Shading-guided manga screening from reference," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[7] Zhansheng Li, Nanxuan Zhao, Zongwei Wu, Yihua Dai, Junle Wang, Yanqing Jing, and Shengfeng He, "Reference-based screentone transfer via pattern correspondence and regularization," in *Computer Graphics Forum*, 2023.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10674–10685.

[9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[10] Lvmin Zhang, Chengze Li, Tien-Tsin Wong, Yi Ji, and Chunping Liu, "Two-stage sketch colorization," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–14, 2018.

[11] Changqing Zou, Haoran Mo, Chengying Gao, Ruofei Du, and Hongbo Fu, "Language-based colorization of scene sketches," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–16, 2019.

[12] Lvmin Zhang, Chengze Li, Edgar Simo-Serra, Yi Ji, Tien-Tsin Wong, and Chunping Liu, "User-guided line art flat filling with split filling mechanism," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9889–9898.

[13] Koki Tsubota, Daiki Ikami, and Kiyoharu Aizawa, "Synthesis of screentone patterns of manga characters," in *2019 IEEE international symposium on multimedia (ISM)*. IEEE, 2019, pp. 212–2123.

[14] Zixin Zhu, Xuelu Feng, Dongdong Chen, Jianmin Bao, Le Wang, Yinpeng Chen, Lu Yuan, and Gang Hua, "Designing a better asymmetric vqgan for stablediffusion," 2023.

[15] Azuma Fujimoto, Toru Ogawa, Kazuyoshi Yamamoto, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa, "Manga109 dataset and creation of metadata," in *Proceedings of the 1st International Workshop on CoMics ANalysis, Processing and Understanding*, 2016.

[16] Minshan Xie, Menghan Xia, and Tien-Tsin Wong, "Exploiting aliasing for manga restoration," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13405–13414.

[17] Anonymous, Danbooru community, and Gwern Branwen, "Danbooru2021: A large-scale crowdsourced and tagged anime illustration dataset," `https://gwern.net/danbooru2021`, 2022.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[19] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[21] Minshan Xie, Chengze Li, Xueting Liu, and Tien-Tsin Wong, "Manga filling style conversion with screentone variational autoencoder," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–15, 2020.

[22] Lvmin Zhang, Xinrui Wang, Qingnan Fan, Yi Ji, and Chunping Liu, "Generating manga from illustrations via mimicking manga creation workflow," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.