ISSN 2616-5775 Vol. 8, Issue 5: 86-93, DOI: 10.25236/AJCIS.2025.080510

Comic Image Generation Based on Diffusion Models

Yanling Zhang

Management Science and Information Engineering, Hebei University of Economics and Business, Shijiazhuang, China, 050061

Abstract: *With the rapid development of the digital entertainment industry, the demand for efficient comic image generation is increasing, especially in areas such as animation production, game design, and personalized content creation. High-quality and automated generation techniques have become essential. However, traditional comic image generation methods rely on manual drawing or basic image processing, making it difficult to achieve both rich details and automation, thus limiting creative flexibility and productivity. To address these challenges, this paper proposes a novel comic image generation model based on diffusion models—Hand-Drawn Comic Diffusion Model (HD-CDM). By learning complex image distributions, HD-CDM progressively refines noisy images to generate comics with intricate line work, vibrant colors, and distinctive artistic styles, significantly improving both quality and efficiency while reducing reliance on manual labor and computational resources. Furthermore, this paper constructs a diverse comic-style image dataset, providing a solid foundation for model training and evaluation, thereby advancing research in this field. Experimental results demonstrate that compared to existing comic image generation models, HD-CDM achieves superior performance in terms of image realism, stylistic consistency, and creative diversity, offering a novel solution for automated comic image generation. In the field of comic creation, it helps artists to quickly generate sketches or concept drawings; for non-professional comic creators, it lowers the threshold of comic creation. They can enter a simple text description or select a specific art style to quickly generate comic images with professional standards, thus making it easier for them to participate in the creation of comics.*

Keywords: *Diffusion, Controlnet, Image Generation, Comics*

1. Introduction

Image generation models have been increasingly recognized in the field of hand-drawn comic creation. The ability to generate high-quality, customized comic images not only reduces production costs but also enhances the efficiency of creators and optimizes the visual experience for readers, opening up new possibilities for comic creation. However, existing comic image generation methods, such as autoregressive models, Generative Adversarial Networks (GANs), and Denoising Diffusion Probabilistic Models (DDPMs), still have certain limitations in terms of image quality, style consistency, and sampling efficiency. For instance, Song Tao proposed Rough Set-DDPM, which integrates rough set theory with DDPM to improve the efficiency of generating high-resolution (HR) images from low-resolution (LR) inputs [1]; Ke Aihua developed SketchDiffusion, a conditionally stable diffusion model for text-guided hand-drawn sketch synthesis, significantly improving controllability and performance [2]; Li Gehui introduced Diff-ReColor, a novel diffusion-based image coloring framework that addresses the challenge of coloring grayscale images with high semantic fidelity and diversity [3]. While these methods have made notable progress in specific tasks, there is still room for improvement in comic-style generation, structural fidelity, and efficient inference.

To address these challenges, this paper proposes a diffusion-based comic image generation method—Hand-Drawn Comic Diffusion Model (HD-CDM)—designed to enhance the quality, style consistency, and creative flexibility of comic image generation. HD-CDM leverages an advanced diffusion model framework to iteratively denoise and refine input images while preserving key characteristics of comic art, such as line artistry, color layering, and structural integrity. Additionally, to support model training and evaluation, this paper has meticulously constructed a high-quality dataset encompassing diverse comic styles, ensuring that the generated comics achieve superior stylistic diversity and visual coherence.

Experimental results demonstrate that, HD-CDM achieves significant improvements in image clarity, style stability, and creative diversity. Through this study, HD-CDM not only introduces a novel research direction for automatic comic generation but also expands the possibilities for hand-drawn comic creation and industry applications. The specific contribution points are as follows:

(1) This paper presents a high-quality classified medium-sized comic dataset, addressing the lack of existing comic datasets.

(2) This paper proposes the HD-CDM model, which integrates an image compression module and a conditional information modeling module, effectively improving control over generated images while significantly reducing computational resource consumption.

(3) This paper introduces the CCM module, which incorporates a ControlNet tailored for comic styles, ensuring that the generated images align more closely with the desired content.

2. Related Work

2.1 Previous Models

Previous models employed GAN-based approaches to generate comics. Zhao Xiuzhi proposed a Global-Local Perceptive GAN with Style Attention to apply thematic features for personalized comic generation [4]. Pan Chuanyu utilized a two-stage reconstruction method to generate 3D cartoon faces with detailed textures [5]. Dong Yongsheng constructed a Two-Stage Generative Adversarial Network (TSGAN) for image cartoonization, enhancing the interpretability of the image generation model [6]. Tong Shuxian introduced a novel Weakly Supervised Exaggeration Transfer Network (ETCari) to learn diverse exaggerated comic styles from different artists [7]. The approach in this article enables the model to better capture key elements in hand-drawn comics, increasing the diversity and richness of the generated images.

2.2 Development of Diffusion Models

Diffusion models have undergone numerous evolutions. DDPM employs a model architecture based on the gradual addition of noise and subsequent reverse denoising. Alex Nichol proposed an improved denoising diffusion probabilistic model capable of generating higher-quality samples [8]. Wang Zhendong introduced Diffusion-GAN, which leverages a forward diffusion chain to generate Gaussian mixture distributed instance noise [9]. Robin Rombach presented latent diffusion models, which significantly improve the training and sampling efficiency of denoising diffusion models [10]. This model has inherited the essence of DDPM's gradual noise addition and reverse denoising process. This preserves its advantages in generating high-quality, diverse images.

2.3 Optimization Module: ControlNet

Lvmin Zhang introduced ControlNet, a neural network architecture designed to learn conditional control for large pre-trained text-to-image diffusion models [11]. Zhao Shihao presented Uni-ControlNet, a unified framework that allows for the simultaneous utilization of different local controls and global controls in a flexible and composable manner within a single model [12]. Li Ming proposed the ControlNet++ method, which improves controllable generation by explicitly optimizing pixel-level cyclic consistency between the generated image and conditional controls [13]. This model significantly upgrades the model's capabilities in conditional control and personalized generation.

3. Model

As depicted in Figure 1, the training process of HD-CDM strategically leverages the properties of the Gaussian distribution. Through a progressive forward diffusion process, noise is incrementally added to the original data, followed by a backward denoising step to restore clear comic images. To further refine the generation quality, this model introduces a KL constraint, effectively minimizing the KL divergence to bridge the gap between the generated comic image distribution and the target distribution. Moreover, this model incorporates an innovative cross-attention mechanism, which empowers the model to produce comic images that closely align with textual descriptions, enhancing their semantic coherence and visual fidelity.

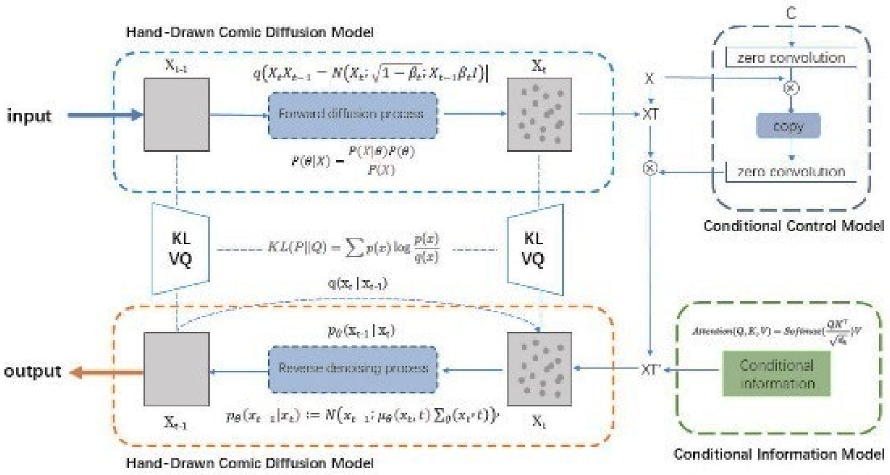
Hand-Drawn Comic Diffusion Model
X_{t-1}
$q(X_t X_{t-1} - N(X_t; \sqrt{1-\beta_t}; X_{t-1}\beta_t I)|$
X_t
input
Forward diffusion process
$P(\theta|X) - \frac{P(X|\theta)P(\theta)}{P(X)}$
X
XT
C
zero convolution
copy
zero convolution
Conditional Control Model
KL
VQ
$KL(P||Q) = \sum p(x) \log \frac{p(x)}{q(x)}$
KL
VQ
$q(x_t | x_{t-1})$
$p_\theta(x_{t-1} | x_t)$
output
Reverse denoising process
XT'
$Attention(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V$
Conditional information
X_{t-1}
$p_\theta(x_{t-1}|x_t) := N(x_{t-1}; \mu_\theta(x_t, t) \sum_\theta(x_t, t))$
X_t
Hand-Drawn Comic Diffusion Model
Conditional Information Model

Figure 1 Flowchart of the Hand-Drawn Comic Diffusion Model

3.1 Dataset

First, this model used a combination of artificial and computer technology to widely collect resources from major forums and websites. By using BeautifulSoup crawler framework and XPath query, we can efficiently capture the rich comic resources on these platforms. In the screening process, this model uses advanced image processing technology to automatically perform preliminary screening of collected images to ensure that the pixel values of the image are within a reasonable range [-255, 255]. Manual review is also an indispensable part. The processing phase focuses on removing watermarks and bad information from the image. This paper integrates Photoshop's capabilities with AI watermarking, ensuring thorough data review and cleanup. In the classification process, this paper divides the comic works into 10 categories, such as ink painting and Japanese comics. Each category contains 2,000 images, forming a medium-sized data set, with a total of more than 20,000 pictures. Finally, the labeling is performed using the WD1.4 labeler. This article generates the corresponding text file (.txt) for each comic image, and record the information contained in the picture in detail.

3.2 Hand-Drawn Comic Diffusion Model (HD-CDM)

3.2.1 Training process of HD-CDM model

During the HD-CDM model training process, this model uses Gaussian distribution for forward diffusion and backward denoising. Through Gaussian distribution, noise, smooth images and extract features can be effectively processed, and the algorithm performance can be improved. A random variable (X) probability density function that obeys the Gaussian distribution is defined as:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{1}$$

Gaussian noise is injected into the data at each step, generating a new latent variable from the latent variable at the previous moment. As the number of steps increases, noise gradually takes up a dominant position, causing the data to gradually change from an ordered state to a completely disordered noise distribution, providing a solid foundation for the subsequent reverse diffusion process. This generation process is defined as a transfer kernel, expressed as:

$$q(X_t|X_{t-1} = N(X_t; \sqrt{1-\beta_t}X_{t-1}, \beta_t I) \tag{2}$$

The reverse denoising process is a key process in data generation, and its goal is to gradually transform pure noise into realistic data samples [14]. This process is achieved by learning a reverse Markov chain, each step of which tries to remove noise from the current noise data and recover the data from the previous time step:

$$p_\theta(x_{t-1}|x_t) := N(x_{t-1}; \mu_\theta(x_t,t), \textstyle\sum_\theta(x_t,t)) \tag{3}$$

The Gaussian distribution provides a clear probability distribution model for each step of latent variables in the reverse denoising process, allowing the model to describe and generate data more accurately. Gaussian distribution has wide applications in image processing, such as Gaussian filtering [15], etc. In hand-painted comic generation, the image can be smoothed with a Gaussian distribution to reduce noise and unnecessary details. The Gaussian distribution can also be used for feature extraction [16], helping the model better identify and understand key elements and features in hand-drawn comics.By adjusting parameters such as the mean and variance of the Gaussian distribution, hand-painted comics with different styles and characteristics can be generated, thereby increasing the diversity and richness of the generated images.

3.2.2 Generation process of HD-CDM model

In the process of image generation, this model uses a probabilistic statistics method based on Bayes theorem [17]. Bayesian formula gives a method of how to update the prior probability distribution after observing the data:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{4}$$

In hand-painted comic generation, there is uncertainty in the generated image because the input description or instructions may be blurry or diverse. Bayesian decisions are able to deal with this uncertainty by updating beliefs, thereby generating images that are more consistent with expectations. When training a diffusion model, this paper uses Bayesian formulas to optimize the parameters and structure of the model to better adapt to the tasks generated by hand-drawn comics. By calculating the posterior probability under different parameters, this article selects the optimal parameter combination to improve the performance and stability of the model. The generation process of hand-drawn comics involves a lot of randomness and uncertainty. Using Bayesian formulas, this paper quantifies and analyses these uncertainties to generate more realistic and delicate hand-painted comic works.

3.2.3 Picture compression module based on conditional constraints

In image processing, this paper uses VQ constraints to represent image data as a limited number of discrete vectors, thereby reducing data complexity and reducing storage requirements. In addition, Kullback-Leibler divergence constraints are used to measure the difference between two probability distributions [18]. The KL divergence can be expressed as:

$$KL(P||Q) = \sum p(x) log \frac{p(x)}{q(x)} \tag{5}$$

In hand-drawn comic generation based on diffusion model, the KL constraint can be used to ensure similarity between the generated comic image distribution and the target comic image distribution. By adding a conditional constraint-based image compression module, the images generated by HD-CDM have high-quality lines, colors and textures.

3.2.4 Conditional information modeling module

Cross-Attention Mechanism is a technology widely used in deep learning [19]. The cross-attention mechanism allows the model to dynamically focus on different regions in the image according to the conditional information. This interactivity allows the model to capture the correspondence between text and image more accurately, thereby generating images that are more in line with text description. In the Stable Diffusion model, queries usually come from conditional encoders, while keys and values come from image features. The specific formula is as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

By introducing a cross-attention mechanism, the model can generate hand-drawn comics more accurately based on text descriptions. Users can customize the style and content of the hand-drawn comic by entering different text descriptions. The cross-attention mechanism enables the model to generate hand-drawn comics with personalized characteristics based on these descriptions.

3.3 Conditional Control Model (CCM)

The Conditional Control model is an end-to-end neural network architecture that controls pre-trained large diffusion models to support additional input conditions. It is able to divide the weights of large diffusion models into "trainable replicas" and "locked replicas", where locked replicas retain network capabilities learned from large numbers of images, while trainable replicas learn by training on task-specific datasets Conditional control. In the field of image generation, simple keyword control methods often cannot meet the needs of precise control of details. The CCM model can accurately guide the diffusion model to generate images through additional input conditions, thereby solving this problem [20].

During training, the CCM will lock the parameters of the original large model so that its ability to keep the original model unforgettable. At the same time, it will copy a copy of the parameters of the original Encoder part of the model as a trainable copy. The two models are partially connected by zero convolution to ensure that harmful noise is not added to the trainable replica at the beginning of training. In the process of hand-drawn comic generation based on diffusion model, the CCM model can accurately and stably control the generation process of hand-drawn comics.

4. Experiment

4.1 Ablation experiment

4.1.1 Experimental Purpose

This ablation experiment is designed to evaluate the performance of different model configurations on image generation to reveal the impact of individual modules on overall performance. This paper will compare the following four model configurations. (1) Use only the basic picture as input. (2) Based on the basic picture, introduce conditional information for control. (3) Based on the basic picture, use the ControlNet model to generate images. (4) Combine basic pictures, conditional information modeling and ControlNet model for image generation.

4.1.2 Experimental Settings

(1) Dataset: As mentioned in 3.1, this paper built a medium-sized dataset focusing on different comic styles, including more than 20,000 images, covering various styles.

(2) Evaluation indicators:

a) Peak signal-to-noise ratio (PSNR): Based on MSE, but in units of decibels (dB), higher values indicate better image quality [21]. b) Mean Square Error (MSE): Calculate the average value of the sum of squares of the difference between the generated image and the target image [22]. c) Structural Similarity Index (SSIM): measures the similarity between two images in brightness, contrast and structure, ranging from -1 to 1, and 1 means exactly the same [23].

4.1.3 Experimental results

As shown in Figure 2, after introducing conditional information modeling, the comics generated by the model are more consistent in style, theme and plot. ControlNet significantly improves the accuracy of generated images by introducing additional conditions, allowing the model to better capture details. After combining basic information, conditional information modeling and ControlNet model, the model's performance is optimal. This shows that when taking into account the three, the model can capture the features and details of the image more accurately.

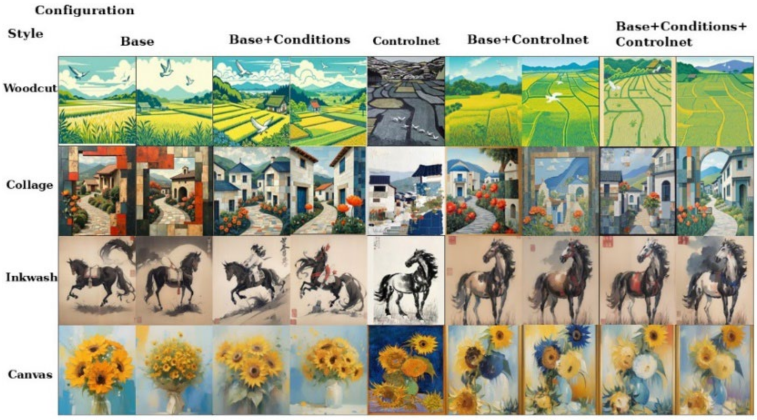

Figure 2 Comparison of ablation experiment

As shown in Table 1, the generated ink-style images are closest to the target image at pixel level and structurally, with the best quality. Woodcut and collage styles are lower in PSNR and SSIM and higher in MSE. This suggests that there is a difference between the generated image and the target image, but these styles show unique artistic effects that suit certain creative scenarios. The oil painting style SSIM is negative, indicating that the generated image is very different from the target image in structure, but some unique visual effects are generated, suitable for experimental creation and stylized exploration.

Table 1 Comparison of evaluation metrics between generated images and dataset images

Comparison	PSNR(dB)	MSE	SSIM
Woodcut	7.918	10501.881	0.111
Collage	7.850	10667.755	0.067
Inkwash	11.881	4216.773	0.643
Canvas	8.052	10182.049	-0.296

4.2 Comparison with existing models

In order to verify the performance and advantages of HD-CDM, this paper conducts a comparative experiment with several other mainstream image generation models.

4.2.1 Experimental Settings

(1) Model selection: a) HD-CDM: the hand-drawn comic-style image generation model proposed in this article. b) DEADiff: an image generation model based on deep learning, focusing on high-quality image generation and detail restoration [24]. c) Kolors: An innovative image generation model focusing on color management and stylized generation. d) HiDiffusion: an efficient image generation model, focusing on the rapid generation of high-quality images [25].

(2) Evaluation indicators: Style consistency: Evaluate the consistency of the generated image and the style of hand-painted comics. Detail richness: measures the performance of generated images in details. Subjective evaluation: Through questionnaire surveys, human subjective feelings and evaluations of generated images are collected.

4.2.2 Experimental results

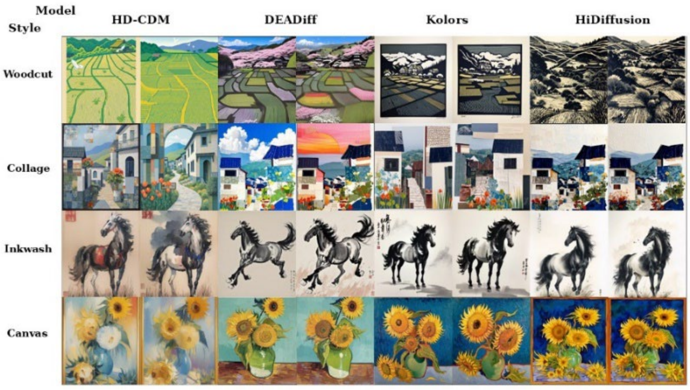

Figure 3 Comparative experimental results diagram

As shown in Figure 3, DEADiff has excellent performance in color management and stylized generation, but its style is more inclined to general artistic creation. HD-CDM focuses on comic style, supports black and white, grayscale and full-color comic generation. Kolors performs excellently in color details, but is not as good as HD-CDM in line and structural details. HD-CDM can generate clearer lines and richer picture levels. HiDiffusion has the advantage in generation speed, but its high-resolution generation capability is limited. In summary, compared with these models, HD-CDM focuses more on the field of comic creation and can better meet the diverse needs from personal creation to commercial comic production.

4.2.3 User subjective evaluation

This paper conducted a user survey to collect subjective evaluations of the HD-CDM model. Each participant received a questionnaire containing five independent questions. For each question, this article randomly selected a comic image from the test set and generate a hand-drawn comic using DEADiff, Kolors, HiDiffusion, and HD-CDM models. This article randomly presented the generation results of these models and asked respondents to rate aesthetics, accuracy, and similarity. Respondents were asked to rate each aspect from 1 to 5, where 1 = very poor and 5 = very good.

This paper invited 258 participants (150 males and 108 females) from various backgrounds. Table 2 shows the mean scores and standard deviations for each question and model for all participants. The results show that HD-CDM is a high-performance, highly flexible image generation model designed for comic creation, and is deeply loved by the public.

Table 2 User subjective evaluation

	Aesthetics	Accuracy	Similarity
DEADiff	2.98/1.25	2.34/1.24	2.25/0.92
Kolors	2.24/1.03	2.67/1.21	2.45/1.09
HiDiffusion	2.79/1.36	3.02/1.26	2.96/1.05
HD-CDM	3.32/0.85	3.45/0.98	3.16/0.64

5. Conclusions

In summary, the HD-CDM model has significant advantages. This paper carefully compiled medium-sized high-definition comic datasets are superior to existing datasets in clarity, diversity and user reviews. The feature extraction formula effectively captures the core features of comics and improves the model's performance in comic-style image generation. The integrated HD resolution module uses advanced technology and conditional GAN to achieve fine conversion from low resolution to high-definition comic pictures. Looking ahead, this model plans to further expand the model functions and introduce sentiment analysis modules to enable the model to generate comic pictures that are more in line with emotional expression based on text input. These potential expansions will bring more innovation and possibilities to the hand-painted comics industry.

References

[1] Song T, Wen R, Zhang L. RoughSet-DDPM: An Image Super-Resolution Method Based on Rough set Denoising Diffusion Probability Model[J]. Tehnički vjesnik, 2024, 31(1): 162-170.

[2] Ke Aihua, Huang Y J, Yang J, et al. Text-guided image-to-sketch diffusion models[J]. Knowledge-Based Systems, 2024, 304: 112441.

[3] Li G, Zhao S, Zhao T. Diff-ReColor: Rethinking image colorization with a generative diffusion model[J]. Knowledge-Based Systems, 2024, 300: 112133.

[4] Zhao X, Chen W, Xie W, et al. Style attention based global-local aware GAN for personalized facial caricature generation[J]. Frontiers in Neuroscience, 2023, 17: 1136416.

[5] Chuanyu P A N, Guowei Y, Taijiang M U, et al. Generating animatable 3D cartoon faces from single portraits[J]. Virtual Reality & Intelligent Hardware, 2024, 6(4): 292-307.

[6] Dong Y, Li L, Zheng L. TSGAN: A two-stage interpretable learning method for image cartoonization[J]. Neurocomputing, 2024, 596: 127864.

[7] Tong S, Liu H, He Y, et al. Weakly Supervised Exaggeration Transfer for Caricature Generation With Cross-Modal Knowledge Distillation[J]. IEEE Computer Graphics and Applications, 2024.

[8] Nichol A Q, Dhariwal P. Improved denoising diffusion probabilistic models[C]//International conference on machine learning. PMLR, 2021: 8162-8171.

[9] Wang Z, Zheng H, He P, et al. Diffusion-gan: Training gans with diffusion[J]. arxiv preprint arxiv:2206.02262, 2022.

[10] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695.

[11] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 3836-3847.

[12] Zhao S, Chen D, Chen Y C, et al. Uni-controlnet: All-in-one control to text-to-image diffusion models[J]. Advances in Neural Information Processing Systems, 2024, 36.

[13] Li M, Yang T, Kuang H, et al. ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback: Project Page: liming-ai. github. io/ControlNet_Plus_Plus[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 129-147.

[14] Chung H, Lee E S, Ye J C. MR image denoising and super-resolution using regularized reverse diffusion[J]. IEEE transactions on medical imaging, 2022, 42(4): 922-934.

[15] Nandan D, Kanungo J, Mahajan A. An error-efficient Gaussian filter for image processing by using the expanded operand decomposition logarithm multiplication[J]. Journal of ambient

intelligence and humanized computing, 2024, 15(1): 1045-1052.

[16] He Y, Bai W, Wang L, et al. SOH estimation for lithium-ion batteries: An improved GPR optimization method based on the developed feature extraction[J]. Journal of Energy Storage, 2024, 83: 110678.

[17] Chun P, Kikuta T. Self-training with Bayesian neural networks and spatial priors for unsupervised domain adaptation in crack segmentation[J]. Computer-Aided Civil and Infrastructure Engineering, 2024, 39(17): 2642-2661.

[18] Chuanyu P A N, Guowei Y, Taijiang M U, et al. Generating animatable 3D cartoon faces from single portraits[J]. Virtual Reality & Intelligent Hardware, 2024, 6(4): 292-307.

[19] Li H, Wu X J. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach[J]. Information Fusion, 2024, 103: 102147.

[20] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 3836-3847.

[21] Martini M G. Measuring objective image and video quality: On the relationship between SSIM and PSNR for DCT-based compressed images[J]. IEEE Transactions on Instrumentation and Measurement, 2025.

[22] Al Najjar Y. Comparative analysis of image quality assessment metrics: MSE, PSNR, SSIM and FSIM[J]. International Journal of Science and Research (IJSR), 2024, 13(3): 110-114.

[23] Martini M. A simple relationship between SSIM and PSNR for DCT-based compressed images and video: SSIM as content-aware PSNR[C]//2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2023: 1-5.

[24] Qi T, Fang S, Wu Y, et al. Deadiff: An efficient stylization diffusion model with disentangled representations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 8693-8702.

[25] Zhang S, Chen Z, Zhao Z, et al. Hidiffusion: Unlocking higher-resolution creativity and efficiency in pretrained diffusion models[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 145-161.