Structure-Aware Latent Diffusion for High-Quality Line Art Colorization

Shuhua Xu[1], Qiang Ai[2], An Zhao[3], Guan Yang[4]*, Bo Chen[5]
School of Intelligent Clothing and Apparel, Zhongyuan University of Technology, Zhengzhou, China[1]
School of Information Science and Engineering, Lanzhou University, Lanzhou, China[2]
School of Computer Science, Zhongyuan University of Technology, Zhengzhou, China[3]
School of Intelligent Perception and Instrumentation, Zhongyuan University of Technology, Zhengzhou, China[4]
School of Mathematical Sciences, Shenzhen University, Shenzhen, China[5]
Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen, China[5]

***Abstract*—To address the limitations of existing line art colorization methods in structural preservation, color mapping accuracy, and semantic consistency, this study proposes a structure-aware multi-instance constrained line art colorization method based on latent diffusion. Built upon the latent diffusion framework, the proposed method introduces a structure-aware constraint mechanism to enhance the preservation of line contours and edge details during generation. Meanwhile, instance-level semantic modeling and feature fusion strategies are incorporated to achieve coherent local color representation and optimize global semantic consistency. In addition, a unified optimization objective is constructed by jointly integrating structural constraints, color consistency constraints, and regularization terms, thereby improving the visual quality and naturalness of the generated results through collaborative multi-constraint learning. Experimental results on public datasets demonstrate that the proposed method outperforms comparative approaches in terms of FID, PSNR, SSIM, and LPIPS, producing high-quality colorization results with clear structures, natural colors, and strong semantic consistency, which verifies its effectiveness and superiority.**

***Keywords*—Line art colorization; latent diffusion; structure-aware modeling; instance-level semantic modeling; feature fusion; image generation**

I. Introduction

In modern electronic measurement and intelligent vision systems, high-precision image acquisition and processing play a crucial role in automated inspection, analysis, and visualization tasks [1], [2]. In applications such as electronic component detection, industrial visual monitoring, and intelligent measurement instruments, the accurate interpretation of image information directly affects the reliability and precision of measurement results [3], [4]. In recent years, with the widespread adoption of animation-style visual content and line art representations in education, electronic displays, digital media, and intelligent human–computer interaction systems, efficient automatic line art colorization has gradually emerged as an important research topic in image processing and electronic information systems [5], [6].

Traditional manual colorization methods are labor-intensive and time-consuming, and they often suffer from inconsistent color representation and structural ambiguity when handling multiple object instances [7], [8], [9]. These limitations can significantly degrade visualization quality and affect observation accuracy in practical applications [10]. With the rapid development of deep generative models, diffusion models have demonstrated remarkable performance in image synthesis and restoration tasks owing to their strong generative capability and latent-space representation learning ability [5], [11]. Nevertheless, existing diffusion-based line art colorization methods still face several challenges in multi-instance scenarios, including insufficient instance-level semantic understanding, inadequate exploitation of structural contour information, and excessive reliance on complex conditional guidance strategies in latent space [12], [13].

To address the above issues, this study proposes a structure-constrained multi-instance line art colorization framework based on latent diffusion, termed SC-MILC (*Structure-Constrained Multi-Instance Line Art Colorization*). Specifically, the proposed framework consists of three key components: a latent generative backbone (LGB), a structure-aware guidance module (SGM), and an instance semantic modeling and feature fusion module (ISMFF). The latent generative backbone performs efficient color generation in latent space, while the structure-aware guidance module enhances contour preservation and structural consistency by exploiting line art edge information. Furthermore, the instance semantic modeling and feature fusion module enables accurate semantic representation and adaptive feature interaction among multiple instances. On this basis, a unified optimization objective is constructed to jointly model diffusion reconstruction loss, structural consistency constraints, color consistency supervision, and instance fusion regularization, thereby achieving high-quality and controllable multi-instance line art colorization.

The main contributions of this work are summarized as follows:

- A latent diffusion-based generation framework for multi-instance line art colorization is proposed, which improves generation stability and computational efficiency.
- A structure-aware guidance module is introduced to enhance structural consistency and edge fidelity by effectively utilizing line art contour information.
- An instance semantic modeling and gated feature fu-

*Corresponding author

sion mechanism is designed to achieve accurate multi-instance color control while balancing local detail representation and global semantic consistency.

- A unified optimization objective function is developed to provide an effective loss evaluation strategy for latent-space multi-instance line art colorization.

II. Problem Definition

The objective of multi-instance line art colorization is to transform an input line art image into a high-quality color image while achieving accurate color control under multiple object instances. Let the input line art image be defined as:

$$S \in \mathbb{R}^{H \times W}, \tag{1}$$

where, H and W denote the height and width of the image, respectively. The reference instance set is represented as:

$$\mathcal{R} = \{I_i\}_{i=1}^{N}, \quad I_i \in \mathbb{R}^{H \times W \times 3}, \tag{2}$$

where, N denotes the number of object instances and each reference image I_i provides the target color information for the corresponding instance. The spatial mask set associated with the reference instances is defined as:

$$\mathcal{M} = \{M_i\}_{i=1}^{N}, \quad M_i \in \{0,1\}^{H \times W}, \tag{3}$$

where, M_i is a binary mask indicating the spatial region occupied by the i-th instance. The goal of the proposed framework is to learn a conditional mapping function:

$$\mathcal{F} : (S, \mathcal{R}, \mathcal{M}) \mapsto \hat{X}, \quad \hat{X} \in \mathbb{R}^{H \times W \times 3}, \tag{4}$$

such that the generated image $\hat{X}$ satisfies the following constraints.

A. Structural Consistency

The generated result should preserve the contour structures and geometric layout of the input line art image, which can be formulated as:

$$\mathcal{F}_{edge}(\hat{X}) \approx \mathcal{F}_{edge}(S), \tag{5}$$

where, $\mathcal{F}_{edge}(\cdot)$ denotes the edge extraction operator.

B. Instance-Level Color Consistency

For each instance region M_i, the generated colors should remain consistent with those of the corresponding reference instance I_i, which is expressed as:

$$\mathrm{Pool}_{M_i}(\hat{X}) \approx \mathrm{Pool}_{M_i}(I_i), \tag{6}$$

where, $\mathrm{Pool}_{M_i}(\cdot)$ represents the average pooling or feature extraction operation within the masked region M_i.

C. Multi-Instance Semantic Consistency

Under multi-instance conditions, the framework should maintain a balance between local instance-level representations and global semantic information, thereby ensuring natural color distributions and avoiding color conflicts among different instances.

Based on the above constraints, the entire task can be formulated as a conditional generation problem in latent space. Let the latent representation of the input line art image be:

$$z_0 = \mathcal{E}(S), \tag{7}$$

where, $\mathcal{E}(\cdot)$ denotes the encoder. The final generated image is obtained by:

$$\hat{X} = \mathcal{D}\left(\mathcal{F}_{diff}\left(z_0, \{f_i, M_i\}_{i=1}^{N}, F_{struct}\right)\right), \tag{8}$$

where, $\mathcal{D}(\cdot)$ denotes the decoder, $\mathcal{F}_{diff}$ represents the latent diffusion generation backbone, f_i denotes the semantic feature of the i-th instance, and F_{struct} represents the structure-aware feature representation.

Through the above formulation, the multi-instance line art colorization task can be uniformly modeled as an optimization problem under a latent conditional generation framework, jointly constrained by structural consistency, instance-level color supervision, and semantic coherence. This formulation provides a unified theoretical foundation for the subsequent network architecture and optimization strategy design.

III. Methodology

The overall architecture of the proposed SC-MILC framework is illustrated in Fig. 1. The framework consists of four major components: a latent-space generation backbone, a structure-aware guidance branch, an instance semantic modeling module, and a reference feature fusion module. The framework takes a line art image together with multiple reference instance images as conditional inputs and performs progressive denoising generation in latent space, thereby achieving automatic multi-instance colorization and consistent color control.

Specifically, a pre-trained variational autoencoder (VAE) is first employed to project the target image into a low-dimensional latent space, where the diffusion process is conducted to reduce computational complexity and memory consumption. Subsequently, the input line art image is fed into the structure-aware guidance branch to extract contour and edge representations. These structural features are further utilized to construct structure-aware constraints, which enhance spatial consistency during the generation process.

Meanwhile, multiple reference instance images are processed by the semantic encoding module to extract high-level semantic representations. Combined with the corresponding instance region masks, the extracted features form region-aware reference representations with instance-level perception capability. Based on these representations, the reference feature fusion module adaptively integrates global semantic information and local instance features to generate unified conditional control signals.

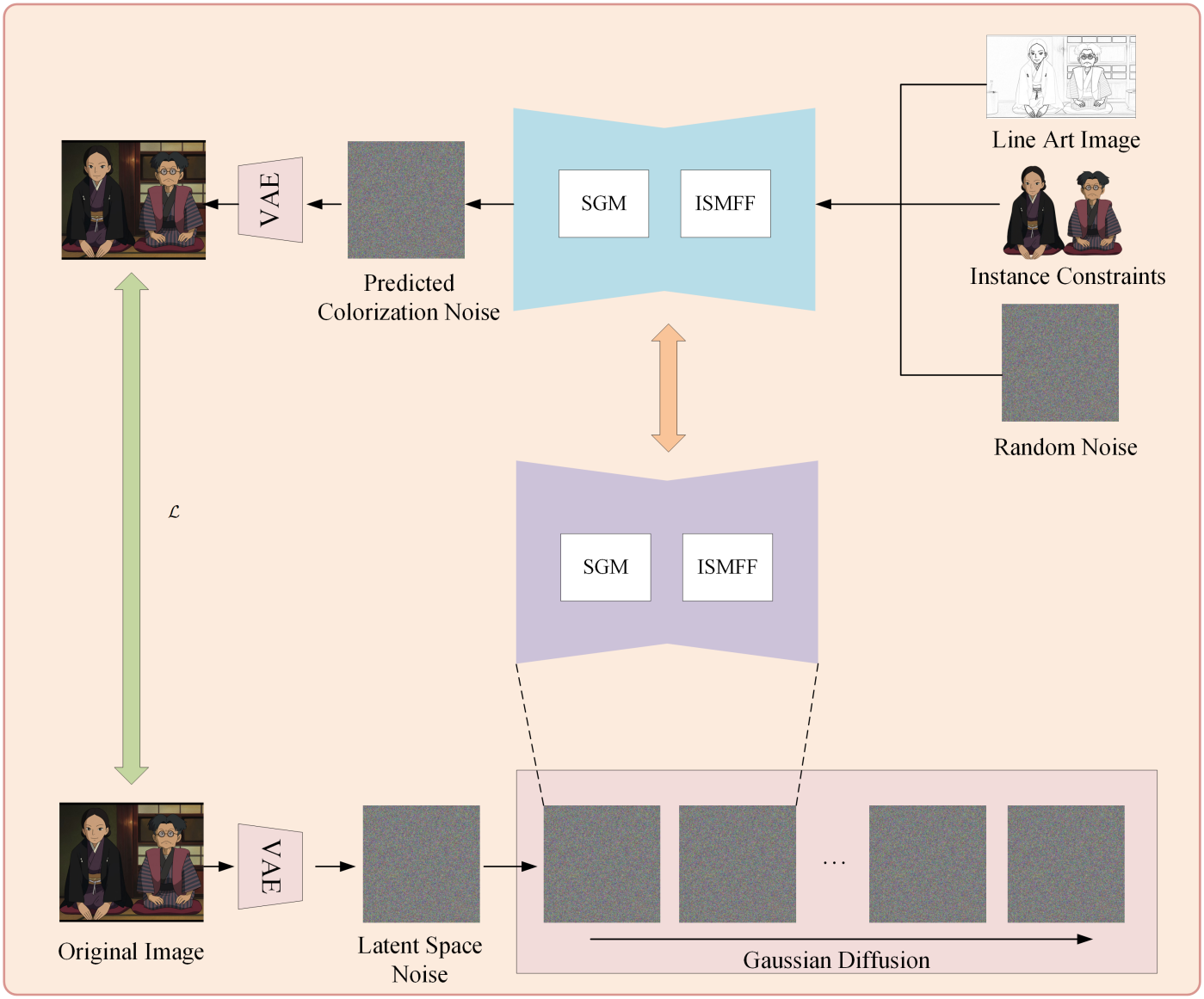

Fig. 1. Overall architecture of the proposed SC-MILC framework. The framework consists of a latent-space diffusion generation backbone, a structure-aware guidance branch, an instance semantic modeling module, and a reference feature fusion module for multi-instance line art colorization.

During the diffusion denoising process, latent features are progressively restored into clear representations under the joint constraints of multiple conditional signals. Among them, structural information constrains the spatial layout and contour consistency of the generated results, while reference semantic information guides color distributions and texture details. Finally, the generated latent representations are decoded into color images through the decoder, producing high-quality multi-instance line art colorization results.

A. Latent Generative Backbone (LGB)

To achieve efficient image generation modeling, a latent-space diffusion generation backbone is constructed in the proposed framework. Let the input image be denoted as $x \in \mathbb{R}^{H\times W\times 3}$. A pre-trained variational autoencoder (VAE) encoder $\mathcal{E}(\cdot)$ is first employed to project the image into a compact latent representation:

$$z_0 = \mathcal{E}(x), \quad z_0 \in \mathbb{R}^{C\times H'\times W'}, \tag{9}$$

where, $H' = \frac{H}{s}$ and $W' = \frac{W}{s}$ denote the latent spatial resolution, and s represents the spatial downsampling factor.

In the latent space, a forward diffusion process is defined by progressively injecting Gaussian noise into z_0 to construct a sequence of latent variables $\{z_t\}_{t=1}^{T}$,

$$q(z_t|z_0) = \mathcal{N}\left(z_t; \sqrt{\bar{\alpha}_t}z_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{10}$$

where, $\alpha_t = 1-\beta_t$, $\beta_t \in (0,1)$ denotes the predefined noise scheduling parameter, and $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$. The above process can be equivalently reformulated as:

$$z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0,\mathbf{I}), \tag{11}$$

which enables direct sampling of noisy latent representations at arbitrary diffusion steps.

Based on the above formulation, a parameterized noise prediction model ϵ_θ is introduced to estimate the injected noise and progressively recover the latent distribution through the reverse diffusion process. The conditional reverse transition probability is defined as:

$$p_\theta(z_{t-1}|z_t, c) = \mathcal{N}\left(z_{t-1}; \mu_\theta(z_t, t, c), \sigma_t^2 \mathbf{I}\right), \tag{12}$$

where, c denotes external conditional information, and the mean parameter μ_θ is estimated by the noise prediction network:

$$\mu_\theta(z_t, t, c) = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t, t, c)\right). \tag{13}$$

Consequently, the latent-space generation process can be uniformly formulated as a learning problem for the noise prediction function $\epsilon_\theta(z_t, t, c)$. The optimization objective is constructed by minimizing the mean squared error between the predicted noise and the ground-truth Gaussian noise:

$$\mathcal{L}_{diff} = \mathbb{E}_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right]. \tag{14}$$

Under the above modeling framework, the latent variables progressively converge from the noise distribution toward the latent representation corresponding to the real data distribution under the guidance of conditional information c. Finally, the generated latent representation is mapped back to the image space through the decoder $\mathcal{D}(\cdot)$:

$$\hat{x} = \mathcal{D}(z_0). \tag{15}$$

The proposed latent generative backbone provides a compact and unified formulation for modeling the transformation from the data distribution to the noise distribution and subsequently to the generated distribution. This formulation further establishes a unified conditional generation foundation for integrating structural guidance and instance-level semantic information in the subsequent modules.

B. Structure-Aware Guidance Module (SGM)

To enhance the structural consistency of the generated results, a structure-aware guidance mechanism is introduced to explicitly incorporate geometric contour information from line art images into the diffusion modeling process. Let the input line art image be denoted as $S \in \mathbb{R}^{H\times W}$. Its structural response is first extracted using an edge operator:

$$E = \mathcal{F}_{edge}(S), \tag{16}$$

where, $\mathcal{F}_{edge}(\cdot)$ denotes the structure extraction operator, and $E \in \mathbb{R}^{H\times W}$ represents the normalized edge intensity map.

To align the structural representation with latent-space features, the extracted edge map is further projected to the latent spatial resolution:

$$E' = \mathcal{R}(E), \quad E' \in \mathbb{R}^{H'\times W'}, \tag{17}$$

where, $\mathcal{R}(\cdot)$ denotes the spatial rescaling operation.

During the diffusion denoising process, let the intermediate latent feature representation be denoted as $F_t \in \mathbb{R}^{C\times H'\times W'}$. The conventional attention operation is formulated as:

$$\mathrm{Attn}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{18}$$

where, Q, K, and V denote the query, key, and value feature mappings, respectively, and d represents the feature dimension.

To introduce structural constraints into the attention mechanism, a structure-aware modulation term A_s is constructed as:

$$A_s = \gamma \cdot \mathrm{Norm}(E'), \tag{19}$$

where, $\mathrm{Norm}(\cdot)$ denotes the normalization operation and γ is a learnable modulation parameter.

The structure-aware attention operation is then formulated as:

$$\mathrm{Attn}_s(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d}} + A_s\right)V, \tag{20}$$

which enables the attention distribution to adaptively focus on structurally important regions.

Furthermore, to strengthen feature representation in edge regions, an element-wise modulation strategy is introduced for intermediate latent features:

$$\tilde{F}_t = F_t \odot (1 + \alpha E'), \tag{21}$$

where, $\odot$ denotes element-wise multiplication and α represents the modulation coefficient. This operation assigns higher response weights to structural edge regions, thereby suppressing color overflow and structural ambiguity during generation.

Through the above mechanisms, structural information is jointly incorporated into the diffusion process via attention bias modulation and feature-level enhancement. Consequently, the proposed structure-aware guidance module enables the diffusion model to preserve clear structural boundaries and stable spatial layouts during latent-space denoising, thereby improving the structural consistency between the input line art image and the generated colorized result.

C. Instance Semantic Modeling and Feature Fusion Module (ISMFF)

To achieve accurate color control under multi-instance conditions, a unified instance semantic modeling and feature fusion mechanism is constructed in the proposed framework. Let the reference image set be denoted as $\mathcal{R} = \{I_i\}_{i=1}^{N}$, and the corresponding instance mask set be represented as $\mathcal{M} = \{M_i\}_{i=1}^{N}$, where $I_i \in \mathbb{R}^{H\times W\times 3}$ and $M_i \in \{0,1\}^{H\times W}$.

First, each reference instance image is processed by a semantic encoder $\mathcal{F}_{enc}(\cdot)$ to extract high-level semantic representations:

$$f_i = \mathcal{F}_{enc}(I_i), \quad f_i \in \mathbb{R}^{C\times H'\times W'}, \tag{22}$$

where, f_i denotes the semantic feature representation of the i-th reference instance.

To establish spatial correspondence between semantic features and instance regions, the instance masks are projected to the latent-space resolution:

$$M_i' = \mathcal{R}(M_i), \quad M_i' \in \mathbb{R}^{H'\times W'}, \tag{23}$$

where, $\mathcal{R}(\cdot)$ denotes the spatial rescaling operation.

Based on the projected masks, instance-aware semantic features are defined as:

$$\hat{f}_i = f_i \odot M_i', \tag{24}$$

where, $\odot$ denotes element-wise multiplication. This operation explicitly constrains semantic features within the corresponding instance regions.

To further enhance global semantic perception, a global semantic representation is introduced as:

$$f_g = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Pool}(f_i), \tag{25}$$

where, $\mathrm{Pool}(\cdot)$ denotes the global pooling operation.

Furthermore, an adaptive feature fusion mechanism is designed to integrate local instance features with global semantic information. The gating function is formulated as:

$$g_i = \sigma\left(W_1 f_g + W_2 \mathrm{Pool}(\hat{f}_i)\right), \tag{26}$$

where, $\sigma(\cdot)$ denotes the Sigmoid activation function, and W_1 and W_2 represent learnable parameters.

Based on the obtained gating weights, the fused instance feature representation is defined as:

$$\tilde{f}_i = g_i \cdot \hat{f}_i + (1 - g_i)\cdot f_g, \tag{27}$$

which enables adaptive balancing between local instance-specific information and global semantic consistency.

Finally, all fused instance features are aggregated to construct a unified instance-aware conditional representation:

$$F_{inst} = \sum_{i=1}^{N} \tilde{f}_i, \tag{28}$$

which is incorporated into the diffusion generation framework as part of the conditional guidance signal:

$$c = \{F_{inst}, F_{struct}\}, \tag{29}$$

where, F_{struct} denotes the structure-aware feature representation generated by the structure-aware guidance module.

Through the above formulation, instance-level semantic alignment is explicitly achieved by introducing mask-based spatial constraints, while the proposed gating mechanism adaptively balances local instance features and global semantic information. Consequently, the proposed ISMFF module enables stable and semantically consistent color mapping in multi-instance scenarios. Compared with conventional methods relying on strict spatial feature alignment, the proposed strategy effectively reduces feature alignment complexity while improving semantic representation capability and model generalization performance.

D. Optimization Objective and Constraint Functions

Under the unified framework of the latent generative backbone, the structure-aware guidance module, and the instance semantic fusion module, an overall optimization objective is constructed to ensure structural consistency and color fidelity in the generated images.

Let $\epsilon_\theta(z_t, t, c)$ denote the predicted noise in latent space and ϵ represent the ground-truth Gaussian noise. The basic diffusion optimization objective is defined as:

$$\mathcal{L}_{diff} = \mathbb{E}_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right]. \tag{30}$$

To enhance structural fidelity, a structure-aware loss based on edge consistency is introduced:

$$\mathcal{L}_{struct} = \mathbb{E}_{z_0}\left[\|\mathcal{F}_{edge}(\mathcal{D}(z_0)) - E\|_2^2\right], \tag{31}$$

where, $\mathcal{D}(\cdot)$ denotes the latent-space decoder, E represents the edge map extracted from the input line art image, and $\mathcal{F}_{edge}(\cdot)$ denotes the edge extraction operator.

To achieve instance-level color consistency, an instance-aware color matching loss is further defined as:

$$\mathcal{L}_{color} = \frac{1}{N}\sum_{i=1}^{N} \|\mathrm{Pool}_{M_i}(\mathcal{D}(z_0)) - \mathrm{Pool}_{M_i}(I_i)\|_2^2, \tag{32}$$

where, $\mathrm{Pool}_{M_i}(\cdot)$ denotes masked average feature extraction within the instance region specified by M_i, and I_i denotes the corresponding reference instance image.

To prevent the gating mechanism from collapsing toward either the local instance branch or the global semantic branch, a gated regularization term is introduced:

$$\mathcal{L}_{gate} = \frac{1}{N}\sum_{i=1}^{N} \|g_i - 0.5\|_2^2, \tag{33}$$

which encourages the gating coefficients to maintain a balanced contribution from both local instance features and global semantic representations during feature fusion. By avoiding extreme gate values, the proposed regularization improves the stability of feature integration and facilitates more effective exploitation of complementary information across different semantic levels.

By combining the above objectives, the final overall optimization function is formulated as:

$$\mathcal{L} = \mathcal{L}_{diff} + \lambda_s \mathcal{L}_{struct} + \lambda_c \mathcal{L}_{color} + \lambda_g \mathcal{L}_{gate}, \tag{34}$$

where, λ_s, λ_c, and λ_g denote balancing hyperparameters used to control the contributions of structural consistency, color supervision, and feature fusion regularization, respectively.

Through the above optimization strategy, the proposed framework enables the diffusion model to simultaneously satisfy the following objectives during latent-space denoising: preserving clear structural contours and spatial layouts from the input line art image, maintaining color consistency between generated instances and reference images, and balancing local instance semantics with global contextual representations to produce visually natural and semantically coherent colorization results.

IV. Dataset and Preprocessing

In this study, two publicly available anime line art datasets, *Sakuga* [14] and *ATD-12K* [15], are adopted to evaluate the performance of the proposed framework on multi-instance line art colorization tasks.

The Sakuga dataset contains high-quality animation frame sequences with rich motion variations and diverse character poses, making it suitable for training and evaluating multi-instance colorization models. In contrast, the ATD-12K dataset provides well-annotated line art and color image pairs covering various character categories and artistic styles, thereby offering reliable references for instance-level color learning.

As illustrated in Fig. 2, both datasets are uniformly processed to construct the final training and testing sets. During preprocessing, samples with excessively low resolution or incomplete line art structures are removed to ensure data quality. For each image pair, the line art image S and the corresponding color reference image I are jointly utilized to establish the conditional colorization relationship. Furthermore, instance masks are extracted as auxiliary conditional inputs for multi-instance semantic modeling.

Specifically, an open-source segmentation tool is first employed to segment characters and key objects in the reference images, thereby generating binary instance masks to support instance-level semantic modeling. Subsequently, both the line art images and the color reference images are uniformly resized to a fixed spatial resolution while preserving the original aspect ratio to maintain latent-space feature alignment.

Fig. 2. Dataset preprocessing and instance mask construction pipeline. The line art images, reference color images, and corresponding instance masks are jointly utilized for multi-instance conditional colorization training.

To improve the generalization capability of the model under varying viewpoints and illumination conditions, data augmentation strategies including random horizontal flipping and brightness perturbation are further applied to the reference images. These augmentation operations effectively enhance the robustness of the proposed framework in modeling instance-level color distributions and structural representations.

The above dataset partitioning and preprocessing strategies ensure the reproducibility of the experiments and provide sufficient multi-instance conditional information for validating the effectiveness of the latent generative backbone, the structure-aware guidance module, and the instance semantic fusion mechanism.

A. Experimental Environment and Parameter Settings

All experiments were implemented using the PyTorch deep learning framework. The detailed experimental environment configuration is summarized in Table I.

TABLE I. Experimental Environment Settings

Environment	Configuration
CPU	Xeon(R) Gold 6348
GPU	NVIDIA A800 80GB
Memory	120 GB
Programming Language	Python 3.9
Deep Learning Framework	PyTorch 2.6.0 + cu118

During the training stage, in order to ensure the stability and convergence of the latent generative backbone and the instance semantic fusion module, the proposed framework was trained for a total of 100,000 optimization iterations. A batch size of 1 was adopted for single-sample training to fully preserve instance-level conditional information during the diffusion learning process.

The initial learning rate was set to 1×10^{-4} and remained fixed throughout the training process to maintain stable latent-space noise prediction and conditional feature optimization. The AdamW optimizer was employed for parameter optimization with momentum coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

In the diffusion process, the total number of diffusion steps was set to $T = 1000$, and a linear noise scheduling strategy was adopted for Gaussian noise injection. The latent-space downsampling factor was set to $s = 8$, which effectively reduced computational complexity while preserving essential semantic and structural information.

For all experiments, the model parameters were initialized using pre-trained latent diffusion weights, and mixed-precision training was adopted to further improve training efficiency and GPU memory utilization.

B. Evaluation Metrics

To comprehensively evaluate the performance of the proposed multi-instance line art colorization framework, four widely used image quality assessment metrics are adopted, including Fréchet Inception Distance (FID), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

FID is employed to measure the distribution discrepancy between generated images and real images in a high-dimensional feature space. A lower FID value indicates that the generated image distribution is closer to the real image distribution, implying higher generation quality and better visual realism.

PSNR is utilized to evaluate the pixel-level similarity between the generated image and the reference image, which is defined as:

$$\mathrm{PSNR} = 10 \cdot \log_{10} \frac{L^2}{\mathrm{MSE}}, \tag{35}$$

where, L denotes the maximum possible pixel intensity value and MSE represents the mean squared error between the generated image and the reference image. Higher PSNR values indicate better reconstruction fidelity and lower pixel-level distortion.

SSIM is adopted to measure structural consistency between generated images and reference images in terms of luminance, contrast, and structural information. It is formulated as:

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{36}$$

where, μ_x and μ_y denote the mean intensities of two images, σ_x^2 and σ_y^2 represent the variances, σ_{xy} denotes the covariance, and C_1 and C_2 are stability constants. An SSIM value closer to 1 indicates better structural preservation and visual consistency.

LPIPS is introduced to evaluate perceptual similarity between generated images and reference images based on deep feature representations extracted from neural networks. It is defined as:

$$\mathrm{LPIPS} = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot (\hat{y}^l_{hw} - \hat{x}^l_{hw}) \right\|_2^2, \tag{37}$$

where, l denotes different network layers, $\hat{x}^l$ and $\hat{y}^l$ represent normalized feature maps, w_l denotes the learned layer weight, and H_l and W_l correspond to the spatial dimensions of the feature maps. Lower LPIPS values indicate higher perceptual similarity and better visual quality.

In summary, lower FID and LPIPS values together with higher PSNR and SSIM values indicate that the generated images are closer to the reference images in terms of distribution consistency, pixel-level reconstruction accuracy, structural preservation, and perceptual quality. These metrics collectively provide a comprehensive evaluation of the proposed method for multi-instance line art colorization tasks.

C. Comparative Experimental Results

To validate the effectiveness of the proposed framework, extensive comparisons were conducted against several representative line art colorization methods. In all comparison methods, the complete reference image was directly used as the conditional input, whereas the proposed framework utilized instance-level information extracted from the same reference image for multi-instance conditional guidance.

For GAN-based approaches, RSIC [16] and SGA [17] were selected as representative comparison methods. Both approaches employ generative adversarial networks for line art colorization, while adopting different network architectures and training strategies. In addition, for diffusion-based methods, AnimeDiffusion [18], ColorizeDiffusion [19], and MangaNinja [20] were adopted for comparison. These methods represent recent advances in diffusion-based line art colorization and provide strong baselines for evaluating color consistency, structural fidelity, and detail restoration under multi-instance conditions.

The quantitative comparison results on the Sakuga and ATD-12K datasets are summarized in Table II.

As shown in Table II, the proposed structure-aware multi-instance line art colorization framework achieves significant performance improvements on both the Sakuga and ATD-12K datasets.

Among the GAN-based approaches, RSIC and SGA obtain relatively high FID values and moderate PSNR and SSIM scores, while their LPIPS values remain comparatively large. These results indicate that GAN-based methods still suffer from color inconsistency and structural distortion, especially in multi-instance scenes or regions containing complex color distributions. In many cases, color overflow occurs around instance boundaries, leading to loss of structural details and degraded visual coherence.

Compared with GAN-based methods, diffusion-based approaches demonstrate clear improvements in image quality and structural preservation. AnimeDiffusion, ColorizeDiffusion, and MangaNinja achieve lower FID values ranging from 25.7 to 22.8 on the Sakuga dataset and from 28.5 to 25.3 on the ATD-12K dataset. Meanwhile, PSNR and SSIM values are improved to the ranges of 24.8–25.4 and 0.801–0.817, respectively, while LPIPS values decrease to 0.165–0.176. These results demonstrate the superiority of diffusion models in maintaining global color consistency and structural fidelity. However, existing diffusion-based methods still exhibit

TABLE II. COMPARATIVE EXPERIMENTAL RESULTS ON THE SAKUGA AND ATD-12K DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Methods	Sakuga				ATD-12K			
	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
RSIC	42.5	21.3	0.732	0.241	48.7	20.9	0.715	0.257
SGA	38.9	22.1	0.749	0.227	44.3	21.7	0.731	0.243
AnimeDiffusion	25.7	24.8	0.801	0.176	28.5	24.1	0.788	0.189
ColorizeDiffusion	23.4	25.2	0.813	0.169	26.1	24.6	0.798	0.181
MangaNinja	22.8	25.4	0.817	0.165	25.3	24.8	0.802	0.178
Proposed Method	**15.2**	**27.1**	**0.856**	**0.121**	**17.4**	**26.3**	**0.841**	**0.135**

Fig. 3. Qualitative comparison results of different line art colorization methods. The proposed method achieves superior structural consistency, color fidelity, and instance-level detail restoration compared with existing GAN-based and diffusion-based approaches.

limitations in instance-level color mapping, particularly for small-scale details and complex multi-instance scenarios.

In contrast, the proposed method achieves the best performance across all evaluation metrics on both datasets. On the Sakuga dataset, the proposed framework reduces the FID score to 15.2, while improving PSNR and SSIM to 27.1 and 0.856, respectively, with LPIPS further reduced to 0.121. Similarly, on the ATD-12K dataset, the proposed framework achieves FID, PSNR, SSIM, and LPIPS values of 17.4, 26.3, 0.841, and 0.135, respectively. These results indicate that the proposed framework effectively preserves structural details while achieving more accurate instance-level color mapping.

In particular, the collaborative interaction between the latent generative backbone and the instance semantic fusion module enables the generated images to maintain high perceptual similarity to the reference images, especially in regions containing small-scale instances or complex color distributions. The proposed structure-aware guidance mechanism further improves edge sharpness and spatial consistency during the diffusion denoising process.

Furthermore, the qualitative comparison results shown in Fig. 3 provide intuitive visual evidence of the performance differences among various methods.

GAN-based methods can generate basic color regions; however, they often exhibit noticeable color overflow and inconsistent color distributions around instance boundaries and small-scale regions. For example, colors from clothing or accessories frequently bleed into adjacent regions, resulting in poor global visual consistency and insufficient detail restoration.

Diffusion-based methods achieve more natural global color distributions and better structural preservation compared with GAN-based methods. Nevertheless, they still encounter difficulties in accurately restoring local color details in complex multi-instance scenarios, and color inconsistencies among different instances remain observable in some cases.

By contrast, the proposed framework demonstrates clear visual advantages. Through the integration of the latent generative backbone, structure-aware guidance, and instance semantic fusion mechanisms, the generated images preserve fine line structures while achieving accurate instance-level color mapping. Particularly in multi-instance scenes containing complex color regions and small objects, the proposed method produces natural color distributions, sharp structural boundaries, and highly consistent color representations across instances.

Moreover, the proposed framework effectively suppresses color bleeding and feature interference among neighboring instances, thereby exhibiting stronger controllability and superior detail restoration capability in multi-instance line art colorization tasks.

Overall, both quantitative and qualitative results demonstrate that the proposed framework significantly improves instance-level controllability, structural fidelity, and overall color consistency, highlighting its effectiveness and practical

value for multi-instance line art colorization applications.

D. Ablation Experiments

To verify the effectiveness of each core component and the unified optimization objective in the proposed framework, a series of ablation experiments were conducted on the Sakuga and ATD-12K datasets. The experiments were designed as follows:

- Experiment A: The structure-aware guidance module (SGM) was removed while retaining the latent generative backbone (LGB) and the instance semantic modeling and feature fusion module (ISMFF). This experiment evaluates the contribution of structural guidance to contour preservation and edge fidelity.
- Experiment B: The ISMFF module was removed while retaining the LGB and SGM modules. In this setting, only global reference information was utilized without instance-level semantic modeling and feature fusion. This experiment evaluates the effectiveness of instance-level semantic guidance for precise multi-instance color control.
- Experiment C: The LGB module was removed while retaining SGM and ISMFF. In this case, latent-space diffusion generation was not employed, and the framework directly performed color mapping based on structural and semantic guidance. This experiment evaluates the importance of latent-space generation for overall image quality and generation stability.
- Experiment D: The complete framework was adopted, including LGB, SGM, ISMFF, and the unified optimization objective. This configuration was used to evaluate the collaborative effectiveness of all proposed modules.

As shown in Table III and Table IV, each module contributes significantly to the overall performance of the proposed framework.

On the Sakuga dataset, Experiment A removes the SGM module while retaining LGB and ISMFF. Without structural guidance, the generated images exhibit blurred edge structures and insufficient contour consistency, resulting in an FID score of 28.5, a PSNR of 24.3, an SSIM of 0.790, and an LPIPS value of 0.182. These results demonstrate that structural information plays an important role in preserving edge sharpness and structural fidelity.

Experiment B removes the ISMFF module while retaining LGB and SGM. Although the overall colorization results remain visually reasonable, obvious color deviations appear in multi-instance regions. The obtained FID, PSNR, and SSIM values indicate that instance-level semantic modeling is crucial for precise multi-instance color control and semantic consistency.

Experiment C removes the latent generative backbone and only retains SGM and ISMFF. In this setting, the overall generation capability deteriorates significantly, with the FID score increasing to 36.8 and LPIPS increasing to 0.215, while PSNR and SSIM decrease substantially. These results demonstrate that the latent diffusion generation backbone provides the fundamental generation capability and stability for the proposed framework.

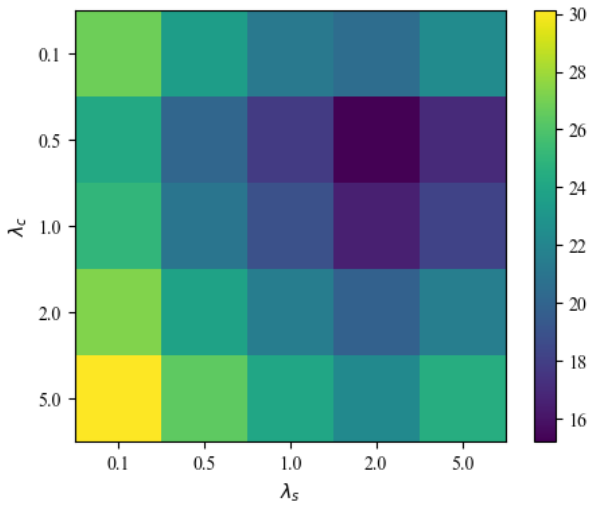

Fig. 4. FID heatmap under different λ_s and λ_c with a fixed λ_g.

Experiment D adopts the complete framework and achieves the best performance across all evaluation metrics. The generated images exhibit clear structural boundaries, accurate color mapping, and high visual consistency, demonstrating the effectiveness of the collaborative interaction among all proposed modules.

Similar trends can also be observed on the ATD-12K dataset. Removing SGM in Experiment A leads to weaker structural preservation and degraded edge quality. Removing ISMFF in Experiment B results in noticeable color inconsistencies among different instances. Meanwhile, removing LGB in Experiment C causes the most severe performance degradation in overall image quality and generation stability. By contrast, the complete framework in Experiment D achieves the best quantitative performance, further validating the effectiveness of the proposed module collaboration strategy.

Overall, the ablation experiments demonstrate that the LGB module provides stable latent-space generation capability and robust color representation, the SGM module enhances structural consistency and edge fidelity, and the ISMFF module enables accurate multi-instance semantic alignment and color control. Moreover, the unified optimization objective further improves high-frequency detail restoration and overall visual quality. These results comprehensively verify the necessity and effectiveness of each proposed component for high-quality and controllable multi-instance line art colorization.

E. Parameter Sensitivity Analysis

To further analyze the coupling relationship among different constraint terms in the proposed optimization objective, a two-dimensional grid search was conducted for λ_s and λ_c while fixing λ_g. The corresponding FID heatmap is illustrated in Fig. 4.

It can be observed that the model performance exhibits a distinct asymmetric distribution in the parameter space, and

TABLE III. ABLATION EXPERIMENTAL RESULTS ON THE SAKUGA DATASET.

Exp.	Modules			Metrics			
	LGB	SGM	ISMFF	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
A	✓		✓	28.5	24.3	0.790	0.182
B	✓	✓		25.1	25.2	0.810	0.170
C		✓	✓	36.8	22.1	0.752	0.215
D	✓	✓	✓	**15.2**	**27.1**	**0.856**	**0.121**

TABLE IV. ABLATION EXPERIMENTAL RESULTS ON THE ATD-12K DATASET.

Exp.	Modules			Metrics			
	LGB	SGM	ISMFF	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
A	✓		✓	31.1	23.8	0.775	0.196
B	✓	✓		27.6	24.7	0.795	0.182
C		✓	✓	39.5	21.7	0.739	0.231
D	✓	✓	✓	**17.4**	**26.3**	**0.841**	**0.135**

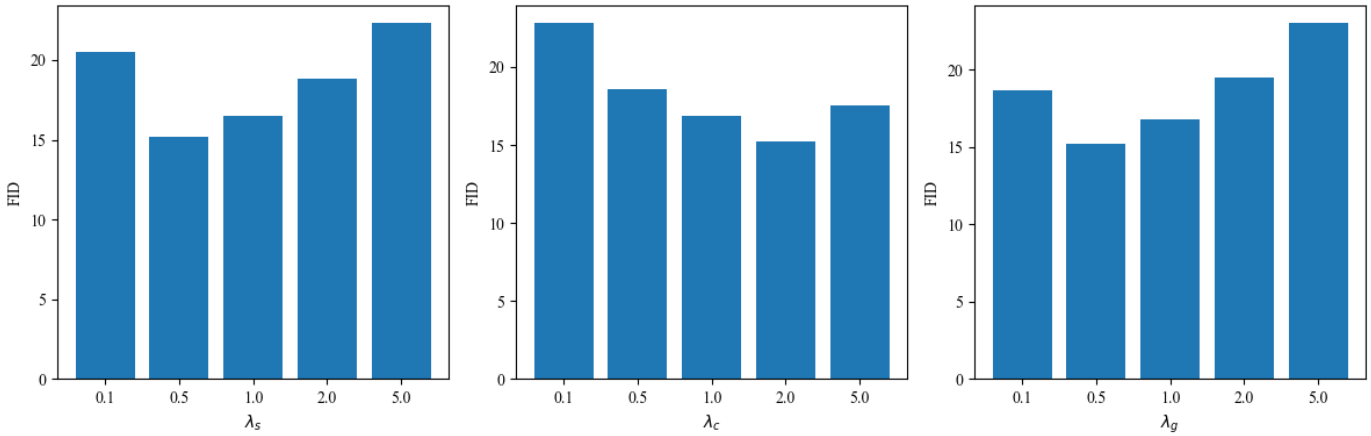

Fig. 5. Results of the univariate sensitivity analysis for different constraint weights.

the optimal performance is achieved around specific non-uniform parameter combinations rather than simple equal-weight settings. This phenomenon indicates that different constraint terms contribute unequally during the optimization process.

Specifically, moderately increasing λ_c helps improve color consistency, enabling the generated images to better match the reference instances. In contrast, λ_s should be controlled within a relatively small range to avoid excessive structural constraints that may suppress detail flexibility. When λ_s becomes overly large, the model excessively emphasizes sketch structures and weakens color representation capability. Conversely, excessively large λ_c values tend to produce overly concentrated color distributions, thereby reducing visual naturalness.

These results demonstrate that a significant trade-off exists between structural preservation and color consistency. Meanwhile, the instance fusion regularization term λ_g achieves the best performance within a moderate range, effectively balancing local instance features and global semantic representations. Overall, the three constraint terms do not contribute equally but instead cooperate through non-uniform weighting strategies, thereby validating the necessity and rationality of the proposed unified optimization objective.

To further investigate the influence of each constraint weight individually, univariate sensitivity experiments were conducted for λ_s, λ_c, and λ_g, while fixing the remaining parameters to their optimal settings. The experimental results are shown in Fig. 5.

It can be observed that all three parameters exhibit a similar trend in which the performance first improves and then deteriorates as the parameter value increases. However, the optimal ranges differ significantly among the three parameters.

The optimal value of λ_s is relatively small, indicating that moderate structural constraints are beneficial for preserving edge clarity, whereas excessive structural guidance suppresses fine-detail generation. In contrast, the optimal value of λ_c is comparatively larger, demonstrating that color consistency plays a more important role in multi-instance line art colorization tasks. Meanwhile, λ_g achieves the best performance within a moderate range. Excessively large fusion regularization tends to over-smooth semantic features and weakens local detail representation.

These results further confirm that different constraint terms play distinct roles during optimization. Therefore, the three constraints should be collaboratively adjusted through non-uniform weighting in order to achieve an effective balance

among structural fidelity, color consistency, and semantic fusion capability.

V. Conclusion and Future Work

This study proposes a structure-aware line art colorization framework based on latent diffusion to address the challenges of insufficient structural preservation, inaccurate color mapping, and semantic inconsistency in line art colorization tasks. The proposed framework introduces a structure-aware constraint mechanism to enhance contour fidelity and spatial consistency during the diffusion generation process. Meanwhile, an instance semantic modeling and feature fusion strategy is designed to improve local color consistency and instance-level controllability. Furthermore, a unified optimization objective is constructed to collaboratively optimize structural constraints, color consistency, and semantic fusion regularization within a unified latent-space generation framework.

Experimental results on publicly available datasets demonstrate that the proposed method achieves superior performance compared with existing approaches in terms of FID, PSNR, SSIM, and LPIPS metrics. The generated colorization results exhibit clearer structural details, more natural color distributions, and stronger semantic consistency. In particular, the proposed framework effectively improves color controllability and detail restoration capability in complex multi-instance scenarios.

Overall, the proposed framework provides an effective solution for high-quality and controllable line art colorization, and offers a feasible technical reference for related image generation tasks and electronic measurement vision applications.

In future work, the proposed framework can be further extended in several directions. First, introducing temporal consistency modeling may improve color stability in animation sequence generation tasks. Second, incorporating multimodal conditional guidance, such as text descriptions or style priors, could further enhance controllability and diversity in color generation. In addition, lightweight diffusion architectures and efficient inference strategies may be explored to improve computational efficiency and facilitate practical deployment in real-time intelligent visual systems.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No. 62573298, and by the Guangdong Provincial Key Laboratory under Grant No. 2023B1212060076.

References

[1] Q. Gao, M. Wu, X. Qin, and L. Hua, "Machine vision driven magnetic particle inspection technology: principles, applications and trends," *Measurement Science and Technology*, vol. 37, no. 3, p. 032001, 2026.

[2] P. Wu, X. He, W. Dai, J. Zhou, Y. Shang, Y. Fan, and T. Hu, "A review on research and application of al-based image analysis in the field of computer vision," *IEEE Access*, 2025.

[3] Z. Li, Z. Geng, Z. Kang, W. Chen, and Y. Yang, "Eliminating gradient conflict in reference-based line-art colorization," in *European conference on computer vision*. Springer, 2022, pp. 579–596.

[4] H. Carrillo, M. Clément, A. Bugeau, and E. Simo-Serra, "Diffusart: Enhancing line art colorization with conditional diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3486–3490.

[5] X. Li, Y. Ren, X. Jin, C. Lan, X. Wang, W. Zeng, X. Wang, and Z. Chen, "Diffusion models for image restoration and enhancement: A comprehensive survey," *International Journal of Computer Vision*, vol. 133, no. 11, pp. 8078–8108, 2025.

[6] Y. Zhang, Y. Ma, B. Wang, Q. Chen, and Z. Wang, "Magiccolor: Multi-instance sketch colorization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 15 205–15 217.

[7] C. Li, "Semantically aware style-controlled animation line art colorization using conditional gans with gcn and attention mechanisms," *Informatica*, vol. 49, no. 29, 2025.

[8] Y. Cao, X. Duan, X. Meng, P. Mok, P. Li, and T.-Y. Lee, "Computer-aided colorization state-of-the-science: A survey," *IEEE Transactions on Visualization and Computer Graphics*, 2025.

[9] J.-X. Chen, L. Lo, S.-Y. Lu, L. Zou, W.-H. Cheng, J. Huh, and S. Lee, "Seco: Semantic-guided multimodal color splash effects," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2026.

[10] J. K. Pandya, S. S. Khandelwal, R. K. Tipu, and K. S. Pandya, "Advancing water quality management: an integrated approach using ensemble machine learning and real-time interactive visualization," *IEEE Access*, 2025.

[11] M. Fuest, P. Ma, M. Gui, J. Schusterbauer, V. T. Hu, and B. Ommer, "Diffusion models and representation learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[12] W. Jiang, Y. Li, Z. Yi, M. Chen, and J. Wang, "Multi-instance imbalance semantic segmentation by instance-dependent attention and adaptive hard instance mining," *Knowledge-Based Systems*, vol. 304, p. 112554, 2024.

[13] S. Fatima, S. Ali, and H.-C. Kim, "A comprehensive review on multiple instance learning," *Electronics*, vol. 12, no. 20, p. 4323, 2023.

[14] Z. Pan, "Sakuga-42m dataset: Scaling up cartoon research," *arXiv preprint arXiv:2405.07425*, 2024.

[15] L. Siyao, S. Zhao, W. Yu, W. Sun, D. Metaxas, C. C. Loy, and Z. Liu, "Deep animation video interpolation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6587–6595.

[16] J. Lee, E. Kim, Y. Lee *et al.*, "Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[17] Z. Li, Z. Geng, Z. Kang *et al.*, "Eliminating gradient conflict in reference-based line-art colorization," in *European Conference on Computer Vision*. Springer, 2022.

[18] Y. Cao, X. Meng, P. Y. Mok *et al.*, "Animediffusion: Anime face line drawing colorization via diffusion models," *arXiv preprint arXiv:2303.11137*, 2023.

[19] D. Yan, L. Yuan, E. Wu *et al.*, "Colorizediffusion: Adjustable sketch colorization with reference image and text," *arXiv preprint arXiv:2401.01456*, 2024.

[20] Z. Liu, K. L. Cheng, X. Chen *et al.*, "Manganinja: Line art colorization with precise reference following," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.