DOI: 10.19884/j.1672-5220.202412012

Expressive Diffusion Network: a Novel Approach to Grayscale Image Colorization Using Diffusion Models

WANG Xingshuo[1,2], WANG Tong[1,2*]
1. College of Information Science and Technology, Donghua University, Shanghai 201620, China
2. Engineering Research Center of Digitized Textile & Apparel Technology, Ministry of Education, Donghua University, Shanghai 201620, China

Abstract: Image colorization has attracted considerable research interest over the past few decades. However, current methodologies frequently struggle with limited local colorization flexibility and produce unnatural color outputs, primarily due to the absence of comprehensive understanding of color perception. In this work, we propose an expressive diffusion network (EDN) that leverages a robust diffusion network to significantly enhance both colorization accuracy and diversity. The EDN consists of two main components: a pre-trained latent diffusion model and a perceptual luminance model based on VQ-Diffusion. These components work together to generate rich and vibrant colors while maintaining high fidelity to the structural features of the original grayscale image. The EDN incorporates controllable creative diffusion (CCD) to direct the color generation process toward more realistic outcomes. Extensive experiments demonstrate that the EDN outperforms existing methods in perceptual quality, offering notable improvements in visual realism and vibrancy across various scenes. The proposed EDN showcases significant improvements over ChromaGAN and InstColor, confirming its robustness in both simple and complex scenarios.

Keywords: diffusion model; image colorization; expressive diffusion network (EDN); controllable creative diffusion (CCD); guided model
CLC number: TP391.4 **Document code**: A
Article ID: 1672-5220(2026)02-0103-09

Open Science Identity (OSID)

0 Introduction

In recent years, generative diffusion models have made significant advancements in image generation tasks, such as image synthesis, restoration, and enhancement. However, effectively colorizing grayscale images remains a challenge, often leading to unsatisfactory results, particularly in preserving the semantic consistency of color and the perceptual richness of generated images. Liu et al.[1] proposed a piggybacked model for cross-task colorization, but its iterative refinement process introduced structural distortions in high-frequency details[2], especially after multiple iterations with blurred or inaccurate images. This challenge is even more prominent in image colorization tasks where it is essential to balance global color richness, visual harmony, and object-level semantics.

While substantial progress has been made in using diffusion-based models for image generation, existing networks still face limitations in controlling the colorization process. The growing importance of diffusion-based text-to-image models has shown their ability to generate highly realistic images by transforming textual descriptions into corresponding visual representations[3]. However, when applied to grayscale image colorization, these models often fall short due to a lack of fine-grained control over the colorization process. Specifically, Fei et al.[4] demonstrated that while text-to-image models could leverage semantic information from text prompts to generate images[5], these models struggled to apply this capability to the nuanced task of colorizing grayscale images. Existing pre-trained models frequently fail to fully leverage grayscale information, resulting in colors that appear muted and lack perceptual coherence. There is thus an urgent need for innovative models that offer detailed guidance and more effective colorization techniques.

In contrast, traditional models such as image transformer and ImageGPT generate images autoregressively by predicting one pixel at a time. These models, while capable of generating images, are computationally expensive and struggle with scalability and resolution, particularly when handling high-resolution images. Other hybrid models like VQ-VAE, VQGAN, and ImageBART aim to reduce image representation complexity by learning discrete latent spaces. Though they are efficient in terms of compression, these models often face training stability issues and may lose detail during the decoding process, resulting in less precise colorization[6].

To address the challenges of traditional diffusion models in grayscale image colorization, this paper

Received date: 2024-12-20
* Correspondence should be addressed to WANG Tong, email: wangtong@dhu.edu.cn
Citation: WANG X S, WANG T. Expressive diffusion network: a novel approach to grayscale image colorization using diffusion models[J]. *Journal of Donghua University (English Edition)*, 2026, 43(2): 103-111.

proposes a novel model based on a diffusion network. By introducing a controllable creative diffusion (CCD) model and VQ-Diffusion decoders, the proposed model enhances control over the colorization process, overcoming the limitations of traditional models and achieving vibrant and consistent colorization.

The contributions of this research are as follows.

1) We propose a novel network based on a pre-trained latent diffusion model to achieve realistic and diverse image colorization and reduce the need for large training datasets and computational resources.

2) We introduce a diffusion-guided model by combining pre-trained weights to produce latent color priors that are closely aligned with the visual semantics of the grayscale input.

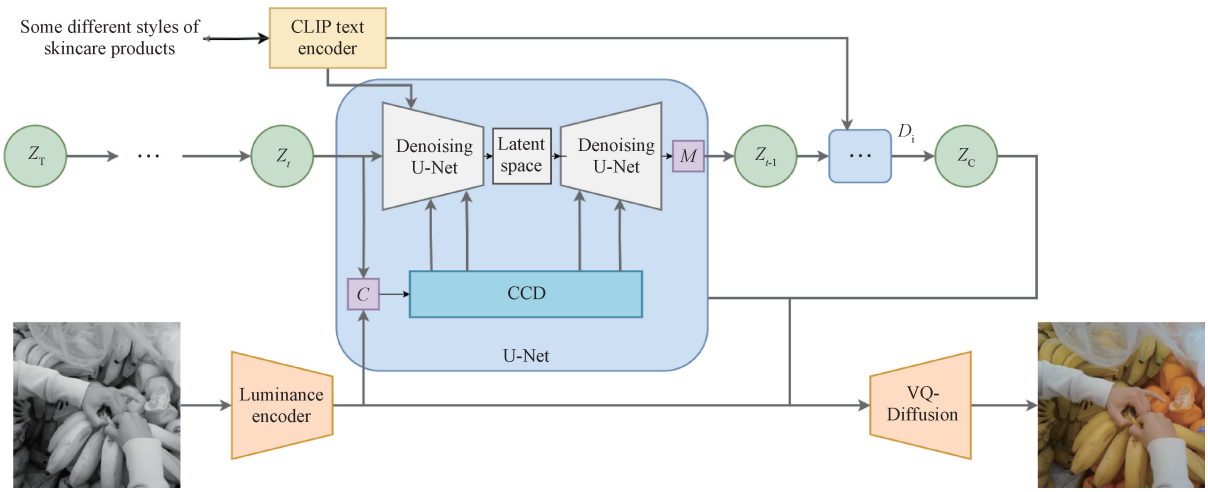
3) We develop a luminance-aware VQ-Diffusion model by using RGB channels to ensure the precise transmission of grayscale information to the decoder for accurate pixel-level coloring.

The remainder of this paper is structured as follows. Section 1 presents the detailed structure of the expressive diffusion network (EDN). Section 2 discusses the experimental results and analyses. Section 3 concludes with a summary of the findings.

1 Methodology

To address the challenges of controlling generated images in traditional diffusion models, we propose an innovative solution that integrates the CCD model within the EDN. The CCD model allows users to control the colorization process by using textual descriptions or color hints, providing high flexibility and customization. In addition, the network incorporates a VQ-Diffusion model, which combines vector quantization with diffusion techniques. By efficiently compressing color information with vector quantization and refining it through the diffusion process, the model achieves improved computational efficiency and visual fidelity, ensuring high-quality and scalable colorization.

As shown in Fig. 1, our network uses a two-step pipeline. First, the CCD model applies a stable diffusion process in the latent space where the user's inputs like text prompts or color hints direct the colorization[8] and allow for precise control. Second, the VQ-Diffusion model refines the compressed color information and merges it with the grayscale input to produce high-quality and vivid results.



CLIP—encoder extracting semantic information from text descriptions; Z_t —intermediate latent variable at time step t in diffusion process; Z_T —initial latent variable representing start of diffusion process; Z_C —transformed latent variable representing final colorized output; M —module for specific processing or transformation steps; C —color information or condition extracted from image; U-Net—architecture used for feature extraction and noise removal; D_i—denoising at step i.

Fig. 1 Architecture of EDN

Regarding the loss function, we combine the losses from the diffusion process, CCD model, and VQ-Diffusion model to achieve both accurate image generation and high-quality colorization. To further optimize the model, we modify the total loss function as

$$L_{\text{total}} = L_{\text{dif}} + \lambda_1 L_{\text{CCD}} + \lambda_2 L_{\text{VQ-Dif}}, \quad (1)$$

where L_{total} is the total loss function to be minimized during training; L_{dif} refers to the loss from the diffusion model, which focuses on predicting noise at each timestep during the reverse diffusion process; L_{CCD} focuses on color consistency and exposure control; $L_{\text{VQ-Dif}}$ drives the model to restore fine details during the reverse diffusion process; λ_1 and λ_2 are the guidance weights of L_{CCD} and $L_{\text{VQ-Dif}}$, respectively. Specifically, we set $\lambda_1 = 0.1$ and $\lambda_2 = 0.6$, which means that the CCD model loss is weighted less than the VQ-Diffusion model loss in the total loss function.

1.1 Diffusion model

The diffusion model is an innovative probabilistic network for learning the data distribution $p(x)$ through a

progressive denoising process. This process can be viewed as the reverse of a Markov chain with a fixed length of T. In traditional networks, image generation in the high-dimensional pixel space directly learns the data distribution, which is computationally expensive and challenging. Our network leverages latent diffusion models to address this challenge by using autoencoders for perceptual compression, allowing us to work in a more efficient low-dimensional latent space. This not only reduces computational complexity but also improves the robustness of the model, enabling better generalization to unseen data.

However, the standard diffusion model still struggles to produce vivid and detailed results, especially in tasks like image colorization. To address this, we incorporate a reweighted version of the variational lower bound, which emphasizes denoising score matching to effectively handle the noise at each denoising step. The motivation is to optimize the generative process during which our model progressively enhances the image's perceptual qualities, allowing us to carefully control the final outcome.

The loss function for the diffusion model in our network is designed to predict the noise added at each step of the diffusion process. It is based on the mean squared error (MSE) between the predicted noise and the actual noise. The loss function is

$$L_{\text{dif}} = \underset{x_0,\, \epsilon,\, t}{E}\left[\left|\epsilon_\theta(x_t,\, t) - \epsilon\right|^2\right], \tag{2}$$

where x_0 represents the original data sample; x_t is the noisy version of the data at timestep t; θ refers to the set of parameters of the model; $\epsilon_\theta(x_t,\, t)$ is the noise predicted by the model at timestep t; ϵ is the actual noise added during the diffusion process, typically sampled from a standard normal distribution. The expectation E is taken over the distribution of the data x_0, noise ϵ, and timestep t.

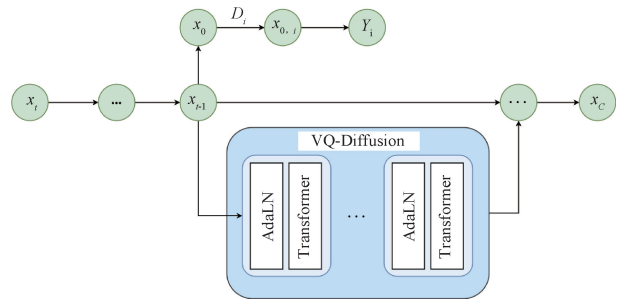
1.2 CCD model

Image colorization[9] is a classic inverse problem in image processing. Traditional networks often face limitations due to the availability of training data and the challenge of generating high-fidelity colorized images from grayscale inputs. We propose the CCD model, which introduces a control mechanism into the diffusion process to enable better colorization results and overcome the constraints of traditional supervised networks, as discussed in Ref. [10].

The key idea behind CCD is to use CLIP's text-image encoder to modulate grayscale image features[11], introducing color hint points to guide the generation process. The use of this model is motivated by the need for control over the colorization process, as traditional diffusion models are often difficult to guide in a targeted way. With CCD, we can inject color hints at each generation step, ensuring that the colorization process is both guided and flexible, aligning more closely with the user intent. In every layer of the generation process, features from the grayscale image I_{g} and the hint point map I_{h} are injected to modulate the diffusion model's feature maps. Referring to the structure in Fig. 2, the specific formula is

$$\widetilde{x}_0 = \frac{x_t}{\sqrt{\overline{\alpha}_t}} - \frac{\sqrt{1-\overline{\alpha}_t}}{\sqrt{\overline{\alpha}_t}}\epsilon_\theta(x_t,\, t), \tag{3}$$

where $\widetilde{x}_0$ represents the original image at the final generation step; α_t denotes the noise schedule parameter for the diffusion process at timestep t; $\overline{\alpha}_t$ is the cumulative product of α_t up to timestep t, and determines the amount of noise removed during denoising. Equation (3) controls how I_{g} and I_{h} modulate the denoising process to guide the colorization.



x_{C} —final colorized output result; $x_{0,\,i}$ —image or intermediate state recovered during decoding process; Y_{i} —intermediate feature information or features extracted from input image.

Fig. 2 Structure of CCD

For the loss function, the addition of color consistency and exposure control losses is critical for ensuring that the generated images are not only accurate in color but also aesthetically pleasing in terms of brightness and overall naturalness[12]. These losses serve to mitigate the common issue of unnatural color distributions that often arise in purely generative models. The complete loss function is

$$L_{\text{CCD}} = \sum_{k=1}^{U}\left|R_k - R_{\text{E}}\right| + \sum_{\forall\,(m,\,n)\,\in\,\epsilon}(Y_m - Y_n)^2 + \sum_{n=1}^{N}\sum_{c\,\in\,\xi}\left(\left|\nabla_{\text{h}}\boldsymbol{M}_{cn}\right| + \left|\nabla_{\text{v}}\boldsymbol{M}_{cn}\right|\right)^2, \tag{4}$$

where the first term ensures global color consistency by minimizing the deviation $|R_k - E|$ between the predicted and reference color distributions; the second term penalizes luminance discrepancies $(Y_m - Y_n)^2$ across adjacent regions (m and n) to maintain smooth exposure transitions; the third term regularizes the gradient magnitude of color maps $\boldsymbol{M}_{cn}$ along horizontal (∇_{h}) and vertical (∇_{v}) directions to reduce artifacts and enforce spatial coherence; n refers to a pixel or position index in the image; k is often used for iterating through color components or layers; U represents the total number; N indicates the number of samples or features; ξ denotes a set or range, possibly for color channels; R_k is the predicted color distribution for component; R_{E} represents

the expected or reference color distribution.

1.3 Luminance -aware VQ-Diffusion model

The problem of detail loss and error accumulation is especially prominent in the latent decoder of VQVAE, as the vector quantization process only approximates the original image's details. Such missing details lead to inaccuracies in the generated image, and structural distortions become more severe in complex scenes. Moreover, VQVAE models face the problem of error accumulation. Since the prediction at each inference step depends on the results of previous steps, early errors progressively accumulate, negatively affecting the quality of subsequent generations. This is particularly problematic in colorization tasks, where the generated colors may not align well with the structure of the grayscale input. The primary cause of this error accumulation is the information loss during VQVAE's quantization process, which limits the model's ability to accurately recover the original details and color distributions.

We propose a new solution by replacing VQVAE with a VQ-Diffusion model. In the VQ-Diffusion model, we improve the latent space representation to avoid the loss of critical details during decoding. The structure of the VQ-Diffusion model is shown in Fig. 2. It gradually restores details during the diffusion process, and thus mitigates distortion and error accumulation. Furthermore, we introduce the grayscale encoder model to enhance the decoding accuracy. The grayscale encoder extracts features from the input grayscale image and directly injects them into the upsampling layers of the VQ-Diffusion decoder. Thus, the decoder leverages both the latent representation and the grayscale features to ensure that the output image aligns perfectly with the original structure. This fusion network effectively enhances pixel-level detail retention and reduces the loss of structural and texture information. The equation for the decoder is

$$F_{\mathrm{Di}'} = F_{\mathrm{Di}} + \mathrm{Conv}_{1\times1}(F_{\mathrm{Gi}}), \tag{5}$$

where $F_{\mathrm{Di}'}$ is the enhanced feature map, and incorporates both the latent representation and the grayscale features; F_{Di} represents the feature map from the VQ-Diffusion decoder at the current stage; F_{Gi} denotes the feature map extracted by the grayscale encoder from the input grayscale image; $\mathrm{Conv}_{1\times1}$ refers to a 1×1 convolutional layer, which is applied to F_{Gi} to ensure that they are in the same dimension as F_{Di}. This fusion process helps the model preserve structural details and improve the overall decoding accuracy.

In the VQ-Diffusion model for grayscale image colorization, we train the model to restore fine details and enhance the image quality by reversing a diffusion process. $L_{\mathrm{VQ\text{-}Dif}}$ is

$$L_{\mathrm{VQ\text{-}Dif}} = L_0 + \sum_{t=1}^{T-1} L_t + L_T, \tag{6}$$

where L_0 is the reconstruction loss at timestep 0 that encourages accurate colorization from tokens, defined as $L_0 = -\log p_\theta(x_0 \mid x_1, y)$; L_t is the KL divergence that quantifies the difference between two probability distributions, guiding model predictions toward the target distribution at intermediate steps and facilitating the reverse diffusion process[13], and $L_t = D_{KL}(q(x'_{t-1} \mid x'_t, x'_0) \parallel p_\theta(x'_{t-1} \mid x'_t, y))$ for $t = 1, 2, \cdots, T-1$; L_T is the final KL divergence at timestep T, and compares the noisy token distribution to the prior, and $L_T = D_{KL}(q(x_T \mid x_0) \parallel p(x_T))$. Here, x'_0 represents the color image tokens; x'_t is the noisy token at timestep t; y is the text condition guiding the colorization; p_θ is the model predicted distribution at each timestep; $q(\cdot, \cdot)$ is the forward diffusion process; $p(x_T)$ is the prior noise distribution at the final timestep. This loss function helps the model recover high-quality color images while progressively reducing noise during the reverse diffusion process.

1.4 User interaction mechanism

The CCD model offers a dual-mode interactive control system, which can be operated either through natural language descriptions or localized color hints, as illustrated in Fig. 3.

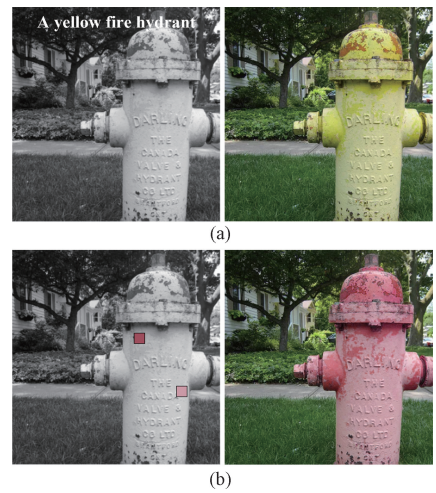


Fig. 3 Example of user interaction mechanism: (a) textual guidance implementation; (b) color hint propagation

When users input semantic directives such as "a yellow fire hydrant" as shown on the left side of Fig. 3(a), the system initially extracts semantic embeddings via the CLIP text encoder. Subsequently, it maps the color constraints to the corresponding object regions through a cross-modal attention mechanism. The fire hydrant is accurately colored in a bright yellow, and maintains a natural gradient effect in non-specified areas as shown on the right side of Fig. 3(a).

For localized fine-grained control (shown on the marked point example on the left side of Fig. 3(b)), after the user marks the target color block in the grayscale image selection area, the CCD model employs an

adaptive propagation algorithm based on color similarity. This mechanism ensures that adjacent pixels maintain color coherence in the hue-saturation lightness (HSL) color space, while strictly adhering to the original gradient features of the input image in the luminance channel, thereby ensuring visual authenticity after color injection.

When no user input is provided, the system automatically invokes the pre-trained VQ-Diffusion codebook, and matches the image semantic content with a historical color scheme library to achieve automated and faithful colorization.

2 Experiments and Analyses

2.1 Datasets

All experiments were conducted on a server equipped with 8 NVIDIA TITAN RTX GPUs, each with 24 GB GDDR6 VRAM, providing substantial computational power and memory capacity for handling large-scale image generation tasks. The experiments were implemented based on PyTorch 1.13, utilizing an open-source latent diffusion network. We employed the pre-trained miniSD and the fine-tuned Stable Diffusion v1.4 model (at an image resolution of 256 × 256 pixels) as the latent diffusion priors for text-to-image generation.

The model training was conducted separately on the COCO-Stuff dataset[14] and ImageNet dataset[15]. For the latent diffusion-guided model, training was performed on the COCO-Stuff dataset, while the luminance-aware VQVAE model was trained on the ImageNet dataset in Ref. [16]. The AdamW optimizer was used for both models with a learning rate r_1 of 0.000 01. To generate color hint maps, we applied the simple linear iterative clustering (SLIC) superpixel segmention and color quantization techniques. During training, 30%–50% of the color hint regions were extracted from the quantized superpixel maps of the original images[17]. For unconditional colorization, the original captions were randomly replaced with generic prompts (such as "colored photo" and "high-quality photo") with 50% probability during training[18].

Our network was evaluated on standard benchmarks, specifically the ImageNet and COCO-Stuff datasets. Following the established protocol, we assessed all networks on a subset of the ImageNet validation set (containing 10 000 images), and conducted comparative tests on the COCO-Stuff validation set (5 000 images).

2.2 Evaluation metrics

The main evaluation criteria for image colorization are perceptual realism and color vividness. Therefore, we adopted the Fréchet inception distance (FID) to measure the distribution similarity between the predicted and ground truth values, which to some extent reflected perceptual realism. To assess color vividness, we employed the chrominance metric, which aligned with human visual perception. Additionally, we reported evaluation results by using the peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS). The lower FID and LPIPS, the better the results; the higher PSNR and SSIM, the better the results. However, it is important to note that plausible colorization results may exhibit significant color differences from ground truth, so these metrics may not accurately reflect actual performance and should be considered as reference points.

2.3 Implementation details

To evaluate the performance of the network in unconditional colorization, we compared it with three categories of state-of-the-art networks that could colorize grayscale images without user input. The first category includes CNN-based networks such as CIColor[19], UGColor[20], Deoldify[21], InstColor[22], ChromaGAN[23], and PiggybackGAN. The second category focuses on transformer-based networks, exemplified by ColTran[24]. The third category is hybrid architectures combining CNNs and transformers, like DISCO[25].

For the task of colorization based on user-provided hint points, we benchmarked our network against two recent advanced networks. These include a CNN-based user-guided network and a transformer-based multimodal colorization network. Our evaluation highlights how these models perform in integrating hint points to achieve accurate and realistic colorization.

To understand the contribution of individual components in the VQ-Diffusion model, we conducted a series of ablation studies. First, we examined the grayscale encoder model, finding that its removal significantly impaired the model's ability to recover the structure and resulted in less precise outputs, particularly in complex scenes. When the grayscale encoder was included, the model preserved local details effectively, ensuring the structural consistency of the generated images with the input. Then, we assessed the mask-and-replace diffusion strategy. Replacing it with a traditional network caused noticeable error accumulation during inference, leading to inaccurate colors. The mask-and-replace diffusion strategy corrected errors and maintained color consistency. Finally, we analyzed the impact of diffusion step counts. Fewer steps accelerated generation but reduced detail and image quality, whereas more steps improved the quality at the expense of slower inference.

The contribution of the CCD model was analyzed by examining color hint points, variance terms, and loss functions. First, the CCD model without hint points exhibited a significant drop in colorization accuracy, underscoring the importance of CLIP-provided hints. Then, the removal of the variance term resulted in improved clarity of the generated images, supporting the hypothesis that eliminating variance enhanced generation performance. Finally, we evaluated the role of the color consistency loss and exposure control loss. The results showed that removing the color consistency loss substantially decreased colorization accuracy, while

eliminating exposure control loss mainly affected the luminance and naturalness of the generated images.

2.4 Experimental results

2.4.1 Quantitative comparison

The results in Table 1 and Table 2 demonstrate the effectiveness of our network on both the ImageNet and COCO-Stuff datasets. On the ImageNet dataset, the proposed EDN achieves an FID of 5.51, slightly outperforming DISCO with an FID of 5.57. On the COCO-Stuff dataset, the proposed EDN obtains the lowest FID of 8.22, outperforming Deoldify with an FID of 12.75. This highlights the effectiveness of our CCD model in reducing FID through controlled color hint points.

Table 1 Results on ImageNet dataset

Network	FID	PSNR/dB	SSIM	LPIPS
CIColor	11.58	21.96	0.897	0.224
UGColor	6.85	**24.26**	0.919	0.174
Deoldify	5.78	23.34	0.907	0.188
InstColor	7.35	22.03	0.909	0.919
ChromaGAN	9.60	22.85	0.876	0.230
ColTran	6.37	21.81	0.892	0.218
DISCO	5.57	20.72	0.862	0.229
EDN	**5.51**	22.78	**0.926**	**0.173**

In terms of the image quality, as shown in Table 2, the proposed EDN achieves a PSNR of 24.44 dB on the COCO-Stuff dataset, an increase of about 7.5% over that of ChromaGAN (22.74 dB). Additionally, SSIM reaches 0.893, surpassing that of ChromaGAN (0.871). The integration of the grayscale encoder in our VQ-Diffusion model aids in preserving structural details, contributing to these enhanced quality metrics.

Table 2 Results on COCO-Stuff dataset

Network	FID	PSNR/dB	SSIM	LPIPS
CIColor	21.44	22.08	0.902	0.217
UGColor	14.74	24.34	**0.924**	0.165
Deoldify	12.75	23.49	0.914	0.181
InstColor	12.24	22.35	0.838	0.238
ChromaGAN	20.57	22.74	0.871	0.233
ColTran	11.65	22.11	0.898	0.210
DISCO	10.59	20.46	0.851	0.236
EDN	**8.22**	**24.44**	0.893	**0.164**

2.4.2 Qualitative comparison

Figure 4 presents visual comparison of various models applied to colorization tasks on different types of images, including simple and complex scenes. In simple scenes, our model generates vivid and realistic colors, effectively avoiding color biases. In complex scenes, the CCD model's mask-and-replace diffusion strategy mitigates color bleeding and produces diverse and natural colorization.

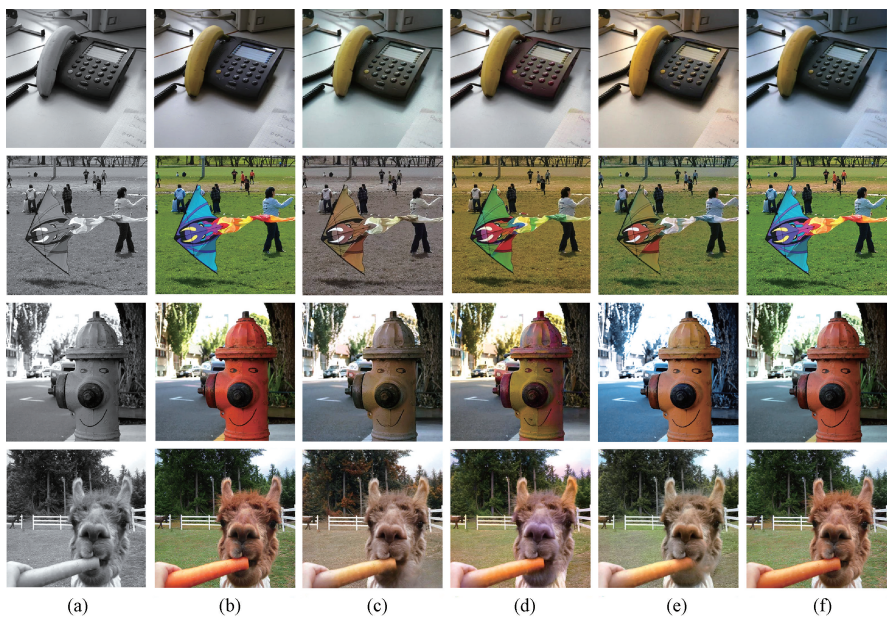


Fig. 4 Qualitative comparison with existing colorization approaches: (a) input; (b) ground truth; (c) ChromaGAN; (d) ColTran; (e) DISCO; (f) EDN

2.5 Ablation studies

The impact of the VQ-Diffusion model was assessed by comparing it with the original KL-f8 VQVAE decoder used in the diffusion model, as well as two fine-tuned versions: ft-EMA and the luminance-aware VQVAE models. The VQ-Diffusion model outperforms the alternatives by leveraging grayscale features, delivering the sharpest results, as illustrated in Fig. 5.

(a) (b) (c) (d) (e)

Fig. 5 Quantitative comparison of reconstructions by different decoders on COCO-Stuff validation set: (a) KL-f8 VQVAE; (b) ft-EMA VQVAE; (c) luminance-aware VQVAE; (d) VQ-Diffusion; (e) ground truth

As summarized in Table 3, our VQ-Diffusion model achieves superior results compared to the baseline KL-f8, ft-EMA, and luminance-aware VQVAE models. Our model reaches a PSNR of 35.6 dB and an SSIM of 0.99, significantly outperforming other versions by utilizing the grayscale encoder to retain local image details and enhance the structural accuracy. Furthermore, the mask-and-replace diffusion strategy reduces error accumulation, contributing to a notable improvement in PSIM, with our model achieving 0.04.

Table 3 Quantitative comparison of reconstructions by different decoders on COCO-Stuff validation set

Decoder	Step	FID	PSNR/dB	SSIM	PSIM
KL-f8 VQVAE	246 000	4.99	23.4	0.69	1.01
ft-EMA VQVAE	313 000	4.42	23.8	0.69	0.96
Luminance-aware VQVAE	44 000	1.88	35.3	0.97	**0.03**
VQ-Diffusion	**42 000**	**1.83**	**35.6**	**0.99**	0.04

The impact of the CCD model was evaluated by guiding the denoising diffusion process toward a subspace aligned with the input grayscale image. As a baseline, we first excluded the proposed model and then applied an alternative guidance mechanism: SDEdit[5]. Additionally, when the CCD model was removed, the model generated less controllable results, as shown in Fig. 6, including images unrelated to the input or distorted shapes of the input image. The results were evaluated by using the default guidance scale and DDIM sampler with 50 steps. For SDEdit, the sampling strength was set to 0.45. Both vanilla-diffusion and SDEdit deviate from the grayscale inputs, while our network delivers superior performance.

(a) (b) (c) (d) (e)

Fig. 6 Quantitative comparison of different diffusion-based image editing networks: (a) input; (b) SDEdit; (c) EDN without CCD; (d) EDN; (e) ground truth

To strengthen the ablation studies, key parameters were systematically analyzed. The temperature parameter, which controls color diversity through VQ-Diffusion codebook sampling, demonstrates stable performance with ± 0.2 variations, with PSNR fluctuations below 0.8 dB within the range of 0.8 to 1.2. The guidance weight proves critical for balancing user inputs and model autonomy: values below 4.0 degrade FID scores by over 33.7% due to insufficient constraint enforcement, and values exceeding 6.0 cause oversaturation artifacts. The denoising step exhibits a quality-speed tradeoff, where 50 steps achieve optimal balance with FID below 30 and the inference latency below 160 ms, consistent with recent diffusion benchmarks for image restoration tasks.

Table 4 Results for parameter sensitivity analysis

Parameter	Range	PSNR/dB	FID
Temperature parameter	0.8–1.2	22.3–23.1	28.2–33.7
Guidance weight	3.0–7.0	20.8–22.9	28.2–42.6
Denoising step	30–70	20.1–23.1	29.8–51.3

2.6 Training and inference efficiency

The proposed EDN integrates the mask-and-replace diffusion strategy of VQ-Diffusion to optimize the training flow. With 8 NVIDIA RTX Titan GPUs configuration, the model achieves full convergence within 64 h (a step of 50 000), showing an efficiency gain of 8%–28% compared to ChromaGAN (70 h) and InstColor (90 h). For inference, the cross-step sampling in VQ-Diffusion enables 300 ms per image latency at a resolution of 512 × 512 pixels (a denoising step of 50), outperforming ChromaGAN (380 ms) and InstColor (450 ms with its three-stage pipeline). This efficiency gain benefits from the VQ-Diffusion-based perceptual luminance model in the backbone, which enables rapid prediction of fine-grained color distributions.

3 Conclusions

In this paper, we introduce an efficient colorization pipeline that leverages pre-trained latent diffusion models for producing realistic and diverse image colorization results. By developing the luminance-aware VQ-Diffusion model that injects grayscale structural details directly into the decoder through quantized latent channels, our network achieves superior precision in both edge preservation and color distribution consistency, thereby addressing the critical challenge of semantic misalignment in diffusion-based colorization. The proposed CCD further harnesses text-to-image priors to generate latent color references aligned with the input's visual semantics, enabling flexible user guidance through textual or chromatic hints. These innovations establish our network as a practical solution for heritage restoration pipelines, where accurate color recovery of archival materials is essential, and for automated colorization tools requiring adaptive control. As diffusion models evolve, networks combining semantic fidelity with granular controllability will remain pivotal for advancing image colorization.

Future directions include extending cross-modal interaction for dynamic user guidance adaptation in heritage restoration, and integrating meta-learning strategies with the VQ-Diffusion backbone to reduce dependency on large training datasets, which is a critical need for rare historical image domains where sample scarcity persists. Theoretical exploration of latent disentanglement and edge-device deployment through quantization remains a critical challenge.

References

[1] LIU H Y, XING J B, XIE M S, et al. Improved diffusion-based image colorization via piggybacked models [EB/OL]. (2023-04-21) [2024-12-11]. https://arxiv.org/abs/2304.11105.

[2] ZHANG W D, ZHUANG P X, SUN H H, et al. Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement [J]. *IEEE Transactions on Image Processing*, 2022, 31: 3997-4010.

[3] LI Y Y, WANG H, JIN Q, et al. SnapFusion: text-to-image diffusion model on mobile devices within two seconds [J]. *Advances in Neural Information Processing Systems*, 2024, 36: 1-12.

[4] FEI B, LYU Z Y, PAN L, et al. Generative diffusion prior for unified image restoration and enhancement [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2023: 9935-9946.

[5] ZHANG Y Z, WU C Y, ZHANG T, et al. Self-attention guidance and multiscale feature fusion-based UAV image object detection [J]. *IEEE Geoscience and Remote Sensing Letters*, 2023, 20: 1-5.

[6] FOSTER D H. Color constancy [J]. *Vision Research*, 2011, 51(7): 674-700.

[7] ZHANG C S, ZHANG C N, ZHANG M C, et al. Text-to-image diffusion models in generative AI: a survey [EB/OL]. (2023-03-14) [2024-12-11]. https://arxiv.org/abs/2303.07909.

[8] LESTER B, AL-RFOU R, CONSTANT N. The power of scale for parameter-efficient prompt tuning [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2021: 3045-3059.

[9] AN Z F, XU Z F, FAN E, et al. Enhancing visual realism: fine-tuning instructPix2Pix for advanced image colorization [EB/OL]. (2023-12-08) [2024-12-11]. https://arxiv.org/abs/2312.04780.

[10] HUANG Y, HUANG J C, LIU Y F, et al. Diffusion model-based image editing: a survey [EB/OL]. (2024-02-27) [2024-12-11]. https://arxiv.org/abs/2402.17525.

[11] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic text-to-image diffusion models with deep language understanding [J]. *Advances in Neural Information Processing Systems*, 2022, 35: 36479-36494.

[12] SONG Y, DHARIWAL P, CHEN M, et al. Consistency models [EB/OL]. (2023-03-02) [2024-12-11]. https://arxiv.org/abs/2303.01469.

[13] 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. New York: IEEE, 2022: 10674-10685.

[14] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context

[C]//ECCV 2014: 13th European Conference. Berlin: Springer, 2014.
[15] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database [C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2009: 248-255.
[16] ZHOU J C, LI B S, ZHANG D H, et al. UGIF-net: an efficient fully guided information flow network for underwater image enhancement [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 1-17.
[17] TANG Z C, GU S Y, BAO J M, et al. Improved vector quantized diffusion models [EB/OL]. (2022-05-31) [2024-12-11]. https://arxiv.org/abs/2205.16007.
[18] ZHANG R, ISOLA P, EFROS A A. Colorful image colorization [C]//ECCV 2016: 14th European Conference. Berlin: Springer, 2016.
[19] GAO D W, FENG Q, WEI Q F, et al. Dyeing of modal fiber in supercritical carbon dioxide using disperse dye CI (color index) disperse yellow 54 [J]. *Journal of Fiber Bioengineering and Informatics*, 2010, 3(3): 148-152.
[20] ZHANG R, ZHU J Y, ISOLA P, et al. Real-time user-guided image colorization with learned deep priors [J]. *ACM Transactions on Graphics*, 2017, 36(4): 1-11.
[21] SALMONA A, BOUZA L, DELON J. DeOldify: a review and implementation of an automatic colorization method [J]. *Image Processing on Line*, 2022, 12: 347-368.
[22] SU J W, CHU H K, HUANG J B. Instance-aware image colorization [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 7965-7974.
[23] VITORIA P, RAAD L, BALLESTER C. ChromaGAN: adversarial picture colorization with semantic class distribution [C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE, 2020: 2434-2443.
[24] KUMAR M, WEISSENBORN D, KALCHBRENNER N. Colorization transformer [EB/OL]. (2021-02-08) [2024-12-11]. https://arxiv.org/abs/2102.04432.
[25] XIA M H, HU W B, WONG T T, et al. Disentangled image colorization via global anchors [J]. *ACM Transactions on Graphics*, 2022, 41(6): 1-13.

表达性扩散网络：一种基于扩散模型的灰度图像着色新方法

王星烁[1,2]，王　彤[1,2*]

1. 东华大学 信息科学与技术学院，上海 201620
2. 东华大学 数字化纺织服装技术教育部工程研究中心，上海 201620

摘　要：图像着色在过去数十年中受到持续关注。然而，因缺乏对颜色感知的深层理解，现有方法的局部着色灵活性受限且色彩输出不自然。在本研究中，我们提出了一种表达性扩散网络(expressive diffusion network，EDN)，该网络利用稳健的扩散网络来显著提升着色精度和多样性。EDN 包含两个核心组件：预训练的潜在扩散模型和基于 VQ-Diffusion 的感知亮度模型。二者协同作用，在保持原始灰度图像结构特征的同时，可生成丰富而鲜艳的色彩。EDN 还引入了可控创意扩散(controllable creative diffusion，CCD)机制，以引导色彩生成趋向更真实的结果。大量实验表明，EDN 在感知质量方面优于现有方法，在多个场景中显著提高了视觉真实感和生动性。该方法在结果上超越了 ChromaGAN 和 InstColor，展示了其在简单和复杂场景中的良好稳健性。

关键词：扩散模型；图像着色；表达性扩散网络(EDN)；可控创意扩散(CCD)；引导模型