ColorControlNet: A Diffusion Model-Based Dual-Branch Framework for Image Colorization

1st Yan Wan*
School of Information and Intelligent Science
Donghua University
Shanghai, China
winniewan@dhu.edu.cn

2nd Xiaojing Wen
School of Information and Intelligent Science
Donghua University
Shanghai, China
2242819@mail.dhu.edu.cn

3rd Li Yao
School of Information and Intelligent Science
Donghua University
Shanghai, China
yaoli@dhu.edu.cn

4th Ru Zhou
Department of General Surgery
RuiJin Hospital LuWan Branch,
Shanghai Jiao Tong University School of Medicine
Shanghai, China
lw10305zr@rjlwh.com.cn

***Abstract*—Image colorization is an essential task in image generation, aimed at producing realistic color results for grayscale images to enhance visual performance. Despite significant progress in deep learning-based methods for natural images, challenges such as uneven color distribution, color bleeding, and structural instability persist in scenes with limited semantic information. To address these issues, ColorControlNet, a dual-branch image colorization framework based on diffusion models, is proposed in this paper. The model takes grayscale images, palette images, and text descriptions as multimodal inputs. The gating mechanism adaptively combines structural features and color priors, enabling precise feature fusion. To efficiently extract structural and color control features, a lightweight encoding network is designed, integrating multi-scale context fusion and axial self-attention mechanisms to enhance local texture representation and consistency modeling across regions. Additionally, an inference-stage enhancement strategy is proposed to suppress local color bleeding and enhance details without increasing training costs. Experimental results demonstrate that the proposed method outperforms existing methods on both a custom flat design dataset and the DomainNet-painting dataset.**

***Index Terms*—diffusion models, gating fusion mechanism, axial self-attention, multiscale feature fusion, discrete wavelet transform**

I. Introduction

Image colorization is an important technology in image generation and content restoration, generating reasonable color distributions based on grayscale structures. It is widely used for tasks such as color transfer in artwork [1], coloring old photographs [2], and restoring historical films [3]. Manual colorization can provide high-quality image coloring, but it heavily relies on the designer's experience and creativity, making the process labor-intensive, inefficient, and costly. In recent years, deep learning-based image colorization methods have made significant progress. However, such methods still face limitations in low-semantic images (such as flat design images). These images often consist of regular color blocks and repeated geometric structures, lacking stable object-level semantics and distinguishable texture cues, making it difficult for the model to form reliable region partitions and boundary perceptions, thereby weakening the effective constraint of color propagation paths. Additionally, when incorporating multiple conditional inputs, such as palettes or text, the constraints between structural information and color priors are often difficult to couple stably. As a result, color constraints may diffuse across boundaries during denoising or generation, leading to uneven color distribution, color bleeding, and structural inconsistency issues. To address these challenges, we propose a dual-branch conditional control framework for image colorization based on diffusion models, called ColorControlNet. This framework separately models structural information and color priors, and uses a gating mechanism in the control branches to adaptively fuse them. After generating unified control features, these features are injected into the diffusion denoising network, effectively coordinating structural preservation and color consistency in low-semantic scenes. A lightweight encoder is designed for efficiently extracting control features, and an inference-stage enhancement module, decoupled from the training process, is proposed for detail compensation and suppression of color bleeding. The main contributions of this paper are:

- The proposal of the ColorControlNet diffusion model-based dual-branch colorization framework, which uses gating fusion to achieve adaptive integration between structural constraints and color priors. In the Stable Diffusion 1.5 setting, the model requires training only 33,299,072 (33.3M) parameters, significantly reducing the trainable parameters by about 91% compared to the ControlNet [4], thus lowering training and deployment costs.
- The design of a lightweight feature encoding network, MulAxialControlNet (MACN), which combines multi-scale context fusion and axial self-attention mechanisms

to improve local texture expression and cross-region consistency modeling, alleviating color unevenness, cross-boundary color bleeding, and structural instability issues in low-semantic image colorization.
- The proposed optional inference-stage enhancement strategy reduces color bleeding and improves high-frequency details without increasing training costs.

II. RELATED WORK

A. Automatic Colorization

Automatic colorization is an image coloring technique that eliminates the need for user intervention. It uses deep neural networks to learn the color distribution in images, enabling fully automated colorization. These methods are primarily divided into two types: methods based on Convolutional Neural Networks (CNN) [5] and methods based on Generative Adversarial Networks (GAN) [6]. CNN-based methods are typically divided into two stages. The first stage involves extracting image features from grayscale images through convolution and pooling. In the second stage, the network uses these features to predict color information through regression [7] or classification [8] tasks, restoring the color image. To improve accuracy, category labels [9], semantic segmentation maps [10], or bounding boxes [11] are often added to the framework. GAN-based methods, on the other hand, use an adversarial training framework with a generator and a discriminator. The generator produces colorized images from grayscale inputs, while the discriminator evaluates their authenticity. Methods like ChromaGAN [12] optimize the colorization process, ensuring that the generated images match real-world color distributions. Additionally, Wu [13] and Kim [9] use pre-trained GAN models to generate color priors, transferring knowledge to improve colorization performance.

B. Conditional Colorization

Conditional colorization methods introduce more user control into the colorization process, allowing for the adjustment of the generated color effects based on specific conditions. Depending on the input type, conditional colorization methods can be categorized into four types: brush-based colorization, text-based colorization, reference-image-based colorization, and palette-based colorization. In the early stages, brush-based colorization relied on similarity measures [14], propagating local color hints across the entire image through spatial shifts or learned by neural networks. Recent methods, such as UniColor [15], use end-to-end frameworks for assisted colorization. Text-based colorization methods guide colorization with natural language descriptions, with feature affine transformations [16] and cyclic attention models [17] helping to fuse text and image features to improve colorization results. Chang et al. [18] focus on matching color information with objects. Reference-image-based colorization methods [19] colorize grayscale images using the color information from reference images. Palette-based colorization methods use predefined color palettes to guide the colorization process, such as the PalGAN [20], which employs a palette-based generative adversarial network for image colorization.

C. Image Generation with Diffusion Models

Diffusion models have made significant progress in the field of image generation, particularly in generating high-quality images and improving generation efficiency. The GLIDE [21] and Imagen models [22] leverage pre-trained visual or language models (such as CLIP [23] and T5 [24]) to encode textual input into latent vectors, significantly enhancing the quality and effectiveness of image generation. To enhance the control capabilities of diffusion models, researchers have proposed several approaches. Some methods utilize input masks [25] to achieve precise control over local regions of an image. Others employ encoded textual features [26] to improve the understanding of text descriptions. Additionally, certain techniques decompose images into multiple independent control components [27], enabling finer-grained control. Diffusion models have been applied to image colorization. L-CAD. [28] utilizes latent diffusion models for image colorization guided by flexible textual descriptions. Qiu et al. [29] combine text and palette information to enhance color consistency in generated images. However, these methods still face challenges in the collaborative modeling of image structure and color information, leading to issues with structural fidelity and color spillover. We propose a dual-branch control framework based on diffusion models, which, through an adaptive gating fusion mechanism, precisely coordinates the fusion of structural information and color priors, maintaining structural fidelity and color consistency in low-semantic scenes.

III. METHODOLOGY

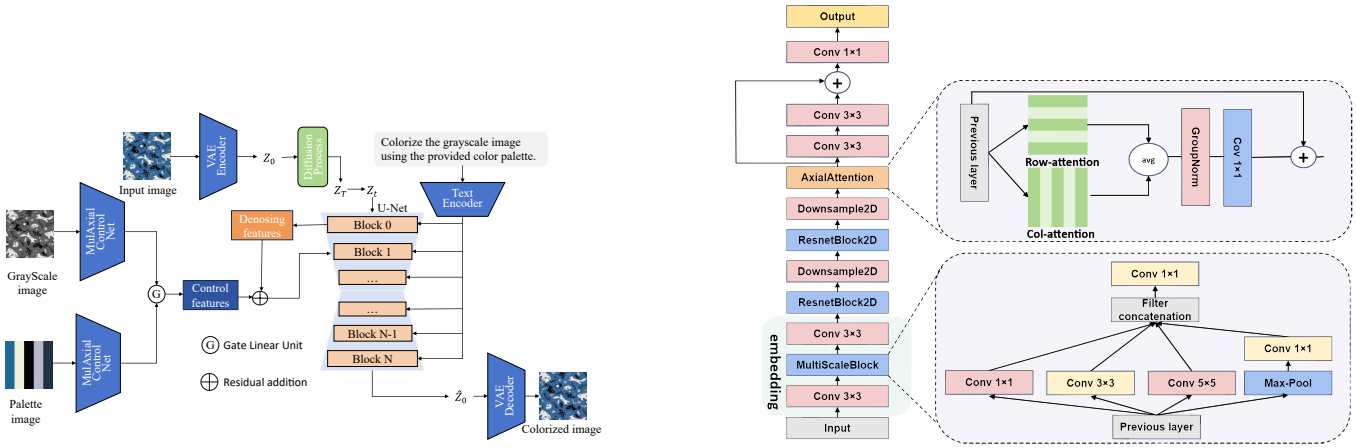
A. Overall Network Architecture

We propose a palette-guided image colorization framework based on diffusion models, named ColorControlNet, as illustrated in Fig. 1 (a), the framework takes a grayscale image I_{grey}, a palette image I_{pal}, and a text prompt P as conditional inputs. The grayscale image provides structural cues, while the palette image supplies explicit chromatic priors. The text prompt offers high-level semantic guidance.

Two lightweight control branches with identical architecture, referred to as MulAxialControlNet (MACN), are employed to extract control features. One branch processes the grayscale image and produces a time-dependent structural feature F^t_{grey}, while the other processes the palette image and produces a time-dependent chromatic feature F^t_{pal}. Both branches are conditioned on the diffusion timestep t through timestep embeddings. The timestep t is shared with the denoising UNet and corresponds to the noisy latent variable Z_t in the diffusion process.

The grayscale and palette features are merged using a gating fusion mechanism. At each timestep t, a gate map $G^t \in [0, 1]$ is predicted from the combined representation of the two features:

$$G^t = \sigma\big(\text{Conv}_{\text{gate}}\big(F^t_{\text{grey}} + F^t_{\text{pal}}\big)\big), \tag{1}$$



(a) Palette-guided diffusion colorization framework.

(b) Architecture of the MulAxialControlNet (MACN).

Fig. 1. **(a)Overall network architecture for palette-guided diffusion colorization.** Two MulAxialControlNet (MACN) branches independently extract structural and chromatic features, which are fused into control features. These features are then injected into the pretrained Stable Diffusion UNet to guide the denoising process in the latent space. **(b) The architecture of the MulAxialControlNet (MACN)**. It is used for extracting control features from grayscale and palette images.

$$F^t_{\text{fused}} = G^t \odot F^t_{\text{grey}} + \left(1 - G^t\right) \odot F^t_{\text{pal}}, \tag{2}$$

where $\sigma(\cdot)$ denotes the sigmoid activation, $\text{Conv}_{\text{gate}}$ is a 1×1 convolution, and $\odot$ represents element-wise multiplication. The fused control feature is defined as

$$C^t_{\text{ctrl}} = F^t_{\text{fused}}. \tag{3}$$

Generation is performed in the VAE latent space using a latent diffusion model. A ground-truth color image x_0 is first encoded into a latent representation $Z_0 = E_{\text{VAE}}(x_0)$. The forward diffusion process progressively injects Gaussian noise to obtain the noisy latent Z_t:

$$Z_t = \sqrt{\bar{\alpha}_t} Z_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I). \tag{4}$$

During reverse diffusion, a conditional denoising network with a UNet backbone predicts the injected noise:

$$\hat{\epsilon} = \epsilon_\theta(Z_t, t, C_t). \tag{5}$$

The conditioning input C_t consists of a text embedding and the fused control feature:

$$C_t = (C_{\text{text}}, C^t_{\text{ctrl}}), \quad C_{\text{text}} = \tau_\xi(P). \tag{6}$$

The text condition modulates UNet features through cross-attention. The control feature C^t_{ctrl} is injected after the first downsampling block of the UNet via residual addition. Directly adding the control feature to the network input was found to cause training instability, likely due to interference between control and denoising functions. The adopted injection strategy preserves the main denoising pathway while enabling effective structural and chromatic guidance.

The estimate of the clean latent is obtained as

$$\hat{Z}_0 = \frac{Z_t - \sqrt{1 - \bar{\alpha}_t}\, \hat{\epsilon}}{\sqrt{\bar{\alpha}_t}}, \tag{7}$$

and the final colorized image is reconstructed by the VAE decoder:

$$\hat{x} = D_{\text{VAE}}(\hat{Z}_0). \tag{8}$$

During training, the parameters of MulAxialControlNet and the linear layers within the attention blocks of the denoising UNet are updated, while the remaining components are kept frozen. Under this training strategy, the total number of trainable parameters is 33.3M, which is significantly lower than the approximately 360M additional parameters introduced by ControlNet [4]. During inference, sampling starts from $Z_T \sim \mathcal{N}(0, I)$ and proceeds iteratively to $t = 0$ under the guidance of the diffusion scheduler.

B. *MulAxialControlNet*

As illustrated in Fig. 1 (b), an image feature encoding network termed MulAxialControlNet (MACN) is designed in this work. The structural control branch and the color prior branch adopt an isomorphic encoder architecture, which extracts diffusion time-dependent control features from the grayscale structural image I_{grey} and the palette image I_{pal}, respectively. These features serve as external constraints for the diffusion denoising network. Given the diffusion timestep t, the feature extraction processes in the two branches are formulated as

$$F^t_{\text{grey}} = \text{MACN}(I_{\text{grey}}, t), \quad F^t_{\text{pal}} = \text{MACN}(I_{\text{pal}}, t). \tag{9}$$

In the shallow encoding stage, the input image is first projected from pixel space to feature space by an embedding module. A single 3×3 convolution with stride 2 is applied to perform initial downsampling and to extract edge and local texture features. Subsequently, a multi-scale feature fusion module $\text{MS}(\cdot)$ is introduced to jointly model structural in-

formation and contextual cues at different scales. This module consists of four parallel branches, defined as

$$\begin{aligned} f_1(x) &= \mathrm{Conv}_{1\times1}(x), \\ f_2(x) &= \mathrm{Conv}_{3\times3}(x), \\ f_3(x) &= \mathrm{Conv}_{5\times5}(x), \\ f_4(x) &= \mathrm{Conv}_{1\times1}(\mathrm{MaxPool}_{3\times3}(x)). \end{aligned} \tag{10}$$

The outputs of all branches are concatenated along the channel dimension as

$$x_{\mathrm{cat}} = \mathrm{Concat}\big(f_1(x), f_2(x), f_3(x), f_4(x)\big), \tag{11}$$

and then fused and aligned by a 1×1 convolution followed by a nonlinear mapping,

$$\mathrm{MS}(x) = \phi(\mathrm{Conv}_{1\times1}(x_{\mathrm{cat}})), \tag{12}$$

where $\phi(\cdot)$ denotes a nonlinear transformation composed of Group Normalization and ReLU activation.

Based on the shallow embedding, MACN adopts a two-stage hierarchical encoding structure to progressively aggregate features. Each stage consists of a ResnetBlock2D followed by a Downsample2D module. The residual learning mechanism preserves structural information across layers, while downsampling gradually enlarges the effective receptive field. The time embedding is obtained as

$$e_t = \mathrm{TimeEmbed}(t), \tag{13}$$

and the hierarchical feature propagation is expressed as

$$x_{l+1} = \mathrm{Down}(\mathrm{ResBlock}(x_l, e_t)), \tag{14}$$

where l denotes the encoding stage index.

To enhance global consistency across spatial regions, an axial self-attention module is introduced at low-resolution feature maps. This module consists of row-wise attention and column-wise attention, which model long-range dependencies along the horizontal and vertical directions, respectively. The fused attention output is projected and added to the input feature via a residual connection, formulated as

$$\mathrm{ASA}(x) = x + \mathrm{Proj}\big(\mathrm{GN}\big(\tfrac{1}{2}\,(y_{\mathrm{row}} + y_{\mathrm{col}})\big)\big), \tag{15}$$

where $\mathrm{Proj}(\cdot)$ denotes a 1×1 convolution and $\mathrm{GN}(\cdot)$ denotes Group Normalization.

In the output stage, a residual refinement block composed of two 3×3 convolutions is applied to further regularize and refine the intermediate features. A final 1×1 convolution is used to project the features to the target dimension, producing the final control features F^t, corresponding to F^t_{grey} and F^t_{pal} for the grayscale and palette branches, respectively. The combination of multi-scale feature fusion and residual propagation improves structural fidelity and suppresses cross-boundary color diffusion. Axial self-attention further strengthens cross-region consistency through long-range dependency modeling, facilitating coordinated color assignment within the same structural regions and enhancing overall structural stability.

C. Inference-Time Enhancement with SAG and DWT

To further suppress local color bleeding and recover fine details, an inference-time enhancement strategy is applied at each diffusion step based on self-attention guidance (SAG) and discrete wavelet transform (DWT). The strategy operates on the current latent estimate $\hat{Z}_0$ and produces a refined latent variable $\tilde{Z}_0$.

Self-Attention Guidance (SAG). SAG leverages self-attention weights inside the UNet to construct a spatial importance mask, which highlights regions that are sensitive to color bleeding. Let A_t denote the self-attention map obtained from the attention block (computed from query Q and key K with a softmax operation). A spatial response map is derived from A_t, resized to the latent resolution, and normalized or thresholded to obtain a mask M_t. A Gaussian blur is applied to the latent variable Z_0 to produce Z_0^{blur}. The guided latent Z_0^{sag} is then formed by replacing masked regions with the blurred latent:

$$Z_0^{\mathrm{sag}} = M_t \odot Z_0^{\mathrm{blur}} + (1 - M_t) \odot Z_0, \tag{16}$$

where $\odot$ denotes element-wise multiplication. A residual pull-back step is used to strengthen the correction:

$$\tilde{Z}_0 = Z_0 + \lambda_{\mathrm{sag}} \cdot (Z_0 - Z_0^{\mathrm{sag}}), \tag{17}$$

where λ_{sag} controls the guidance strength.

Discrete Wavelet Transform (DWT) Detail Enhancement. To enhance high-frequency details while preserving global structure, a 2D Haar wavelet decomposition is applied to the latent variable $\tilde{Z}_0$:

$$(LL, LH, HL, HH) = \mathrm{DWT}(\tilde{Z}_0), \tag{18}$$

where LL is the low-frequency approximation and LH, HL, HH are the high-frequency sub-bands. The enhancement is applied only to high-frequency components while keeping LL unchanged. A gain factor $g \geq 0$ is introduced, and a suppression coefficient $\eta \in (0, 1]$ is used for the diagonal band to reduce noise amplification. A spatial mask $M \in [0, 1]$ (constructed from gradient magnitude or attention responses) is downsampled to the sub-band resolution, denoted as $M_{\downarrow}$. The high-frequency modulation is defined as:

$$\begin{aligned} LH' &= (1 + gM_{\downarrow}) \odot LH, \\ HL' &= (1 + gM_{\downarrow}) \odot HL, \\ HH' &= (1 + \eta g M_{\downarrow}) \odot HH, \end{aligned} \tag{19}$$

followed by inverse wavelet reconstruction:

$$\tilde{Z}_0 = \mathrm{iDWT}(LL, LH', HL', HH'). \tag{20}$$

SAG mainly constrains spatially sensitive regions to mitigate color bleeding, while DWT strengthens edge and texture details through controlled high-frequency enhancement. Their combination provides a lightweight inference-time refinement without introducing additional training cost.

IV. EXPERIMENT AND RESULTS

A. Datasets

We collect a dataset of 10,000 flat design images and apply data augmentation techniques, including translation, flipping, and random cropping, to expand the dataset to 20,000 images. Among them, the last 2,000 images are used as the test set. In addition, 4,000 images from the painting subset of the DomainNet dataset [30] are used for evaluation. For each image, the luminance channel (L) in the Lab color space is extracted as the grayscale input. To obtain palette guidance, K-means clustering is applied to the color channels to generate five representative color values for each image. These colors are then visualized as a 512×512 palette image, which serves as an additional conditional input during training. For quantitative evaluation, we adopt a reference palette setting, where the palette condition is extracted from the corresponding ground-truth color image using K-means clustering (K=5). This setting provides an upper bound under ideal palette constraints, and is used to assess both the colorization quality and the model's adherence to the given palette. In practical applications, the palette condition is typically provided by the user or extracted from a user-provided reference image.

B. Implementation Details

All experiments are conducted based on Stable Diffusion 1.5. During training, the parameters of MulAxialControlNet (MACN) and selected linear layers in the UNet denoising network are updated, while the remaining parameters are kept frozen. The image resolution is set to 512×512, with a batch size of 2. The model is trained for 50 epochs. The learning rate is set to 5×10^{-5} for MACN and 5×10^{-6} for the trainable UNet parameters. During inference, DDIM sampling is employed with 50 denoising steps to ensure stable and high-quality image generation. To further enhance visual quality, self-attention guidance (SAG) and discrete wavelet transform (DWT)-based detail enhancement are applied at inference time. The guidance strength for SAG is set to 0.1, while the high-frequency enhancement strength for DWT is set to 0.05. All baselines were retrained on our flat-design dataset using the same training protocol for fair comparison.

C. Comparative methods

We compare the proposed method with representative colorization approaches. As shown in Fig. 2, BigColor [9] often yields inaccurate colors and distorted textures, DDColor [2] suffers from color bleeding and detail loss, and DISCO [31] produces uneven or incorrect color assignments. In contrast, our method generates more consistent color distributions, preserves fine textures, and effectively mitigates color bleeding while maintaining structural coherence.

Quantitative comparisons are conducted using three widely adopted evaluation metrics: structural similarity index measure (SSIM), peak signal-to-noise ratio (PSNR), and learned perceptual image patch similarity (LPIPS). SSIM evaluates structural similarity, PSNR measures pixel-level reconstruction accuracy, and LPIPS reflects perceptual similarity beyond pixel-wise differences. As shown in Table I, quantitative results on both the Flat-design test set (2,000 images) and the painting subset of the DomainNet dataset [30] (4,000 images) are reported and compared with BigColor [9], DDColor [2], DISCO [31], and PCGDM [29]. On the Flat-design dataset, our method achieves the best performance across all metrics. Similar gains are observed on DomainNet-painting, where it consistently attains higher SSIM and PSNR and lower LPIPS than competing methods.

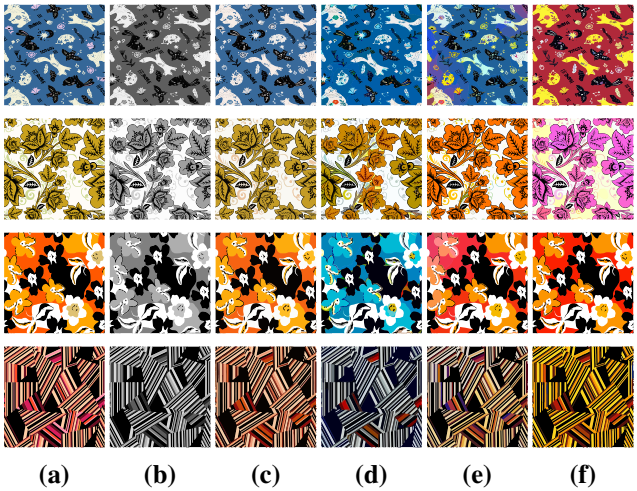


Fig. 2. Overall comparison of different image colorization methods. Columns (a)–(f) correspond to the following: (a) Ground truth color image, (b) Input grayscale image, (c) Result from our method, (d) Result from BigColor, (e) Result from DDColor, and (f) Result from Disco.

TABLE I
QUANTITATIVE COMPARISON ON THE FLAT-DESIGN AND DOMAINNET-PAINTING TEST SETS.

Method	**Flat-design**			**DomainNet-painting**		
	SSIM↑	**PSNR↑**	**LPIPS↓**	**SSIM↑**	**PSNR↑**	**LPIPS↓**
Bigcolor [9]	0.7482	15.43	0.2182	0.7813	18.80	0.1672
DDcolor [2]	0.7621	15.57	0.2179	0.7928	19.48	0.1562
DISCO [31]	0.7377	14.73	0.2364	0.7504	17.26	0.3221
PCGDM [29]	0.7563	14.36	0.2058	0.7736	16.41	0.1946
Ours	**0.8138**	**21.76**	**0.1248**	**0.8030**	**20.96**	**0.1472**

D. Ablation Study

Gating mechanism. To verify the effectiveness of the proposed gating mechanism, an ablation study is conducted by comparing the gated fusion strategy with direct feature fusion. The direct fusion approach simply performs element-wise addition of the grayscale features and the palette features. As shown in Table II (a), the gated fusion consistently outperforms direct fusion across all evaluation metrics. In particular, the SSIM increases from 0.7928 to 0.8138, indicating improved structural preservation, while the LPIPS decreases from 0.1437 to 0.1248, demonstrating a clear enhancement in perceptual quality. The PSNR also shows a slight improvement. These results indicate that adaptive gating plays a critical role in balancing structural fidelity and color consistency.

TABLE II
ABLATION STUDIES.

(a) Gate mechanism.

Model	SSIM↑	PSNR↑	LPIPS↓
Direct add	0.7928	21.63	0.1437
Gate mechanism	**0.8138**	**21.76**	**0.1248**

(b) Palette guidance.

Model	SSIM↑	PSNR↑	LPIPS↓
No Palette	0.8072	16.67	0.2039
Palette	**0.8138**	**21.76**	**0.1248**

(c) Multi-scale and attention in MACN.

MACN Architecture	SSIM↑	PSNR↑	LPIPS↓
No Multi-scale or Attention	0.7530	20.03	0.1626
Only Attention	0.7691	20.55	0.1439
Only Multi-scale	0.7932	21.44	0.1340
Attention + Multi-scale	**0.8138**	**21.76**	**0.1248**

(d) Inference-time strategies.

Configuration	SSIM↑	PSNR↑	LPIPS↓
Baseline (Direct Inference)	0.8138	21.76	0.1248
SAG Guidance	0.8143	21.77	0.1246
DWT Enhancement	0.8174	21.78	0.1247
DWT + SAG Guidance	**0.8178**	**21.80**	**0.1237**

Fig. 3. Colorization results of the same grayscale image using different palettes.

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

Fig. 4. Visual results of ablation on Multi-scale and Axial Attention. Columns (a)–(e) correspond to the following: (a) Input grayscale image, (b) No Multi-scale or Attention, (c) Only Attention, (d) Only Multi-scale, and (e) Attention + Multi-scale.

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

Fig. 5. Visual results of ablation on SAG and DWT. (a) Input image, (b) Baseline, (c) SAG-guided, (d) DWT-guided, and (e) SAG + DWT-guided results.

Effectiveness of palette guidance. Palette input enhances controllability and consistency in colorization by providing explicit color priors. The impact of palette guidance is analyzed on the Flat-design test set. As shown in Table II (b), No Palette refers to the setting without palette guidance. Although SSIM shows a small difference, indicating structural preservation, there are noticeable gaps in PSNR and LPIPS. This suggests that the absence of palette guidance degrades color fidelity and perceptual quality. These results highlight the critical role of palette guidance in improving color consistency and visual quality. Additionally, palette input allows flexible color control, as different palettes applied during inference guide the model to generate diverse color schemes while maintaining clear structural boundaries, as illustrated in Fig. 3.

Multi-scale Feature Fusion and Axial Attention. To evaluate the impact of multi-scale feature fusion and axial attention, an ablation study is conducted with four configurations. As shown in Fig. 4, removing both components degrades detail and color consistency. Axial attention improves structural coherence but still exhibits color bleeding, while multi-scale fusion enhances local details yet causes uneven color distribution. Combining them produces the most stable structures and consistent colors. Table II(c) further indicates that combining both modules yields higher SSIM and PSNR and lower LPIPS, reflecting better structural fidelity and perceptual quality.

SAG and DWT. Fig. 5 and Table II (d) illustrate the effects of introducing SAG and DWT at the inference stage. As shown in Fig. 5, SAG tends to smooth color transitions and alleviate local color bleeding, while DWT mainly enhances high-frequency details such as edges and fine textures. Correspondingly, Table II (d) shows marginal increases in SSIM and PSNR and slight reductions in LPIPS for both strategies. When SAG and DWT are combined, the results exhibit slightly better color consistency and detail preservation, suggesting that the two inference-time strategies provide complementary benefits.

V. CONCLUSION

In this paper, we propose ColorControlNet, a diffusion-based image colorization framework that incorporates multi-modal conditions, including grayscale images, palette images, and text descriptions. A gated fusion mechanism is introduced to coordinate structural constraints and color priors, together with a lightweight MACN encoder for efficient control feature extraction. In addition, inference-time strategies based on self-attention guidance and discrete wavelet transform are explored to refine color transitions and local details. Experimental

results on multiple datasets, particularly flat-design images, indicate that the proposed approach improves color consistency and structural stability compared with existing methods. These findings suggest that explicitly modeling the interaction between structure and color priors can benefit controllable image colorization.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their valuable comments.

REFERENCES

[1] M. Xie, C. Li, X. Liu, and T.-T. Wong, "Manga filling style conversion with screentone variational autoencoder," ACM Trans. Graph., vol. 39, Nov. 2020.

[2] X. Kang, T. Yang, W. Ouyang, P. Ren, L. Li, and X. Xie, "Ddcolor: Towards photo-realistic image colorization via dual decoders," in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 328–338, 2023.

[3] A. Singh, A. Chanani, and H. Karnick, "Video colorization using cnns and keyframes extraction: An application in saving bandwidth," in Computer Vision and Image Processing - 4th International Conference, CVIP 2019, Jaipur, India, September 27-29, 2019, Revised Selected Papers, Part II (N. Nain, S. K. Vipparthi, and B. Raman, eds.), vol. 1148 of Communications in Computer and Information Science, pp. 190–198, Springer, 2019.

[4] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3813–3824, 2023.

[5] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, (USA), p. 415–423, IEEE Computer Society, 2015.

[6] M. Afifi, M. A. Brubaker, and M. S. Brown, "Histogan: Controlling colors of gan-generated and real images via color histograms," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7937–7946, 2021.

[7] J. Zhao, J. Han, L. Shao, and C. G. M. Snoek, "Pixelated semantic colorization," Int. J. Comput. Vision, vol. 128, p. 818–834, Apr. 2020.

[8] Z. Wang, Y. Yu, D. Li, Y. Wan, and M. Li, "Colorful image colorization with classification and asymmetric feature fusion," Sensors, vol. 22, no. 20, 2022.

[9] G. Kim, K. Kang, S. Kim, H. Lee, S. Kim, J. Kim, S.-H. Baek, and S. Cho, "Bigcolor: Colorization using a generative color prior for natural images," in Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII, (Berlin, Heidelberg), p. 350–366, Springer-Verlag, 2022.

[10] A. Shahin Shamsabadi, R. Sánchez-Matilla, and A. Cavallaro, "Colorfool: Semantic adversarial colorization," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1148–1157, 2020.

[11] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware image colorization," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7965–7974, 2020.

[12] P. Vitoria, L. Raad, and C. Ballester, "Chromagan: An adversarial approach for picture colorization," ArXiv, vol. abs/1907.09837, 2019.

[13] Y. Wu, X. Wang, Y. Li, H. Zhang, and X. Zhao, "Towards vivid and diverse image colorization with generative color prior," pp. 14357–14366, 10 2021.

[14] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," in ACM SIGGRAPH 2004 Papers, SIGGRAPH '04, (New York, NY, USA), p. 689–694, Association for Computing Machinery, 2004.

[15] Z. Huang, N. Zhao, and J. Liao, "Unicolor: A unified framework for multi-modal colorization with transformer," ACM Trans. Graph., vol. 41, Nov. 2022.

[16] V. Manjunatha, M. Iyyer, J. Boyd-Graber, and L. Davis, "Learning to color from language," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (M. Walker, H. Ji, and A. Stent, eds.), (New Orleans, Louisiana), pp. 764–769, Association for Computational Linguistics, June 2018.

[17] J. Chen, Y. Shen, J. Gao, J. Liu, and X. Liu, "Language-based image editing with recurrent attentive models," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8721–8729, 2018.

[18] Z. Chang, S. Weng, Y. Li, S. Li, and B. Shi, "L-coder: Language-based colorization with color-object decoupling transformer," in Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII, (Berlin, Heidelberg), p. 360–375, Springer-Verlag, 2022.

[19] Y. Bai, C. Dong, Z. Chai, A. Wang, Z. Xu, and C. Yuan, "Semantic-sparse colorization network for deep exemplar-based colorization," in Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI, (Berlin, Heidelberg), p. 505–521, Springer-Verlag, 2022.

[20] Y. Wang, M. Xia, L. Qi, J. Shao, and Y. Qiao, "Palgan: Image colorization with palette generative adversarial networks," in Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV, (Berlin, Heidelberg), p. 271–288, Springer-Verlag, 2022.

[21] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in Proceedings of the 39th International Conference on Machine Learning (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of Proceedings of Machine Learning Research, pp. 16784–16804, PMLR, 17–23 Jul 2022.

[22] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in Advances in Neural Information Processing Systems (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 36479–36494, Curran Associates, Inc., 2022.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proceedings of the 38th International Conference on Machine Learning (M. Meila and T. Zhang, eds.), vol. 139 of Proceedings of Machine Learning Research, pp. 8748–8763, PMLR, 18–24 Jul 2021.

[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.

[25] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18187–18197, 2021.

[26] A. Hertz, R. Mokady, J. M. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," ArXiv, vol. abs/2208.01626, 2022.

[27] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao, and J. Zhou, "Composer: Creative and controllable image synthesis with composable conditions," in Proceedings of the 40th International Conference on Machine Learning (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds.), vol. 202 of Proceedings of Machine Learning Research, pp. 13753–13773, PMLR, 23–29 Jul 2023.

[28] Z. Chang, S. Weng, P. Zhang, Y. Li, S. Li, and B. Shi, "L-cad: language-based colorization with any-level descriptions using diffusion priors," in *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA, Curran Associates Inc., 2023.

[29] Q. Qiu, J. Mao, and X. Wang, "Exploring palette based color guidance in diffusion models," in Proceedings of the 33rd ACM International Conference on Multimedia, MM '25, (New York, NY, USA), p. 10287–10295, Association for Computing Machinery, 2025.

[30] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1406–1415, 2019.

[31] M. Xia, W. Hu, T.-T. Wong, and J. Wang, "Disentangled image colorization via global anchors," ACM Transactions on Graphics, vol. 41, Nov. 2022. Publisher Copyright: 2022 ACM.