IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 July 2024, accepted 25 July 2024, date of publication 29 July 2024, date of current version 13 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3435485

RESEARCH ARTICLE

Colorize at Will: Harnessing Diffusion Prior for Image Colorization

WANYU YANG, FEIFAN CAI, YANG SHU, ZIHAO ZHANG, QI LIU, AND YOUDONG DING
Shanghai Film Academy, Shanghai University, Shanghai 200072, China

Corresponding author: Youdong Ding (ydding@shu.edu.cn)

ABSTRACT Image colorization, a pivotal aspect of computer vision, employs advanced algorithms to transform grayscale images into realistic colors. This task is inherently challenging due to the need to balance colorfulness and fidelity while preserving local spatial structures and eliminating ghosting effect. To address these issues, Our research introduces a novel pipeline leveraging Stable Diffusion for image colorization, guided by color hint points or textual descriptions. Compared to current text-to-image model, Our key contributions include multi-modal input flexibility, a trainable pixel-level encoder and a controllable feature modulation block. The multi-modal input flexibility allows for the simultaneous use of grayscale images with color hint points and textual descriptions, facilitating the generation of colorized outputs with greater precision and alignment with user instructions. The trainable pixel-level encoder extracts multi-scale features from input images, guiding the diffusion process to capture generative diffusion prior for image colorization, thereby achieving better consistency between the input and output images. Additionally, the controllable feature modulation block is introduced to strike a balance between colorfulness and precision through an adjustable coefficient α. By integrating Stable Diffusion with these innovative guidance advancements, our model overcomes previous limitations and showcases the potential of advanced generative models to produce highly realistic and contextually appropriate colorized images, significantly impacting applications such as historical restoration and contemporary creative processes.

INDEX TERMS Image colorization, diffusion models, color hint points, pixel-level control, multi-modal input.

I. INTRODUCTION

Image colorization, a significant research topic in computer vision, aims to provide natural and realistic colorization to grayscale images using advanced algorithms. Its importance is highlighted by a wide range of applications, including revitalizing historical photos, restoring cinematic footage, enhancing medical imaging, and supporting artistic endeavors. The demand for visually appealing and contextually accurate colored images continues to grow, emphasizing the need for advancements in this field.

Despite its potential in numerous applications, image colorization remains a complex task. Achieving a natural appearance necessitates balancing color saturation and visual coherence while eliminating artifacts such as color bleeding and incomplete colorization.

Early methods [1], [2], [3], [4], [5] relied on convolutional neural networks (CNNs), which often struggled with a lack of semantic understanding, leading to unrealistic colors. Generative Adversarial Networks (GANs) marked a significant improvement by incorporating semantic data, but these methods [6], [7] still faced challenges with detailed images, often resulting in inconsistent colorization and visual artifacts.

Recently, diffusion-based models [8], [9], [10] have demonstrated significant promise in various visual generation tasks, often surpassing GANs in terms of visual quality. These models refine an initial noisy input to generate high-fidelity images, effectively handling complex structures and semantics. Notable Developments like Stable Diffusion

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

[11] and DALLE-2 [12] have set new benchmarks in Artificial Intelligence Generated Content (AIGC), addressing challenges that GANs face with complex images.

Given their potential, models like Controlnet [8] and T2i-Adapter [8] have been proposed to leverage conditional inputs such as canny, sketch, openpose, etc. to guide the diffusion process, underscoring the importance of generative diffusion prior in visual generation tasks. Nevertheless, these methods struggle with achieving pixel-level control, which is crucial for aligning outputs with conditional images.

Unlike general image generation tasks, image colorization typically requires the output to closely match the input image. However, the inherent randomness of pre-trained diffusion models poses challenges for image colorization. To achieve effective colorization, solely fine-tuning is far from sufficient, and retraining these text-to-image (T2I) models is extremely resource-intensive.

To tackle the above difficulties, we propose incorporating pixel-aware conditional control into the diffusion process to obtain authentic and precise image outcomes. Our approach introduces a novel pipeline based on Stable Diffusion, guided by grayscale images with optional color hint points or textual description inputs. Without the need for retraining, our method leverages the advantages of the pre-trained Stable Diffusion model and integrates additional pixel-level inputs, offering a more precise and cost-effective solution to the challenges of colorization tasks. Key contributions of our work include:

- **Enhanced Flexibility with Multi-Modal Input.** Inspired by UniColor [13], We introduce the concept of color hint points, which are specific markers placed on the target grayscale image. Unlike traditional approaches that solely use textual input for Stable Diffusion models, our method supports the simultaneous input of grayscale images with color hint points, and optionally text with color terms. This flexibility improves the model's ability to generate colored outputs that align with user preferences.
- **Pixel-Level Guidence.** Acknowledging the inherent limitations of existing pre-trained models in providing pixel-level control based on images, which often struggle to maintain spatial structures. Our method introduces an additional trainable pixel-level encoder to extract multi-scale features from input images. This design can achieve better consistency between the input and output images.
- **Balance Between Colorfulness and Precision.** Due to the randomness of diffusion process, the output images frequently diverge from the ground truth. To address this issue, We introduce a controllable feature modulation block, which aims to strike a balance between colorfulness and precision through an adjustable coefficient α.

Our architecture showcases the potential of advanced generative models to transform grayscale images into colorized ones, significantly impacting widespread applications such as historical restoration and contemporary creative processes.

II. RELATED WORKS

Image colorization can be categorized into automatic colorization and user-guided colorization based on the involvement of user interaction.

A. AUTOMATIC COLORIZATION

Automatic Colorization primarily employs deep learning methods, leveraging deep neural networks to learn the mapping from grayscale images to colored images. With the emergence of large-scale datasets, these methods can colorize grayscale images using data-driven approaches without requiring user input or reference images. Cheng et al. [1] proposes the first image colorization approach based on an end-to-end deep neural network. In order to obtain diverse colorization results, Deshpande et al. [2] suggested using a Variational Autoencoder (VAE). Su et al. [3] believes that foreground-background separation aids in improving colorization effectiveness, utilizing a detection model to extract bounding boxes as prior.Recently, with the development of Generative Adversarial Networks (GANs), some works [6], [7] have attempted to use pre-trained GAN models for colorizing multiple classes of natural images. Following the success of transformer [14] in natural language processing, transformer have rapidly expanded into the computer vision domain. The Vision Transformer (ViT) [15] has witnessed rapid development in various downstream vision tasks [16], [17], [18], [19]. In the field of image colorization, ColTran [20] is the first to employ transformers to build a probability model for sampling colors from the learned distribution. This model possesses conditional generation capabilities, generating rough low-resolution colorized images, which are subsequently enhanced in quality through upsampling techniques, ultimately producing detailed color high-resolution images. CT2 [21] innovatively treats the colorization task as a classification task. This method inputs image blocks and corresponding color labels into a ViT-based network. The network also includes a luminance selection module containing pre-computed dataset probability distributions for more accurate colorization processing.

B. USER-GUIDED COLORIZATION

User-guided colorization involves user interaction to guide the colorization process.

1) STROKE-BASED COLORIZATION

Methods generally require users to place color markers or lines in specific areas of the input image as color conditions. These colors are then propagated to nearby pixels based on color similarity and diffused to specific boundaries through optimization algorithms. Levin et al. [22] introduced color strokes into optimization algorithms as a linear constraint, leveraging prior knowledge that neighboring pixels have similar colors. The optimization objective is to minimize the difference between the color of each pixel and the weighted average of colors of its neighboring pixels, achieving a more consistent colorization outcome. With the later development of deep neural networks, the approach to feature extraction has gradually shifted from manual design to automatic learning from large-scale datasets. Endo et al. [23] explicitly learns similarity mappings through deep neural networks based on pixel similarity, addressing constrained optimization problems. Recently, Xiao et al. [24] introduces an end-to-end neural network architecture for direct color propagation. Zhang et al. [25] proposes a novel pipeline based on the U-Net architecture, allowing for real-time colorization based on user interaction. To enhance spatial consistency in colorization results, Kumar et al. [20] fully leverages the advantages of transformer, designing a transformer-based colorization model that, with reduced reliance on the quantity of strokes, effectively propagates stroke colors to distant similar regions.

2) EXEMPLAR-BASED COLORIZATION

Exemplar-based colorization methods primarily involve colorizing grayscale images based on reference images provided by users. By computing correspondences between the input and reference images, the colors from the reference image are transferred to the input image. Welsh et al. [26] learn color transfer by matching brightness and texture in pixel neighborhoods. Later, methods at different levels are proposed, including pixel-level [27], [28], segmentation region-level [29], [30], [31], and superpixel-level [32], [33], [34]. These methods perform exceptionally well when the original grayscale image and the reference image share the similar content. However, the challenge lies in finding a suitable reference image, which is often time-consuming and lacking spatial semantic consistency. With the advancements in deep neural networks, Lu et al. [35] addresses these challenges by designing a network based on a gated attention mechanism. This network effectively fuses semantic colors from the reference image and global color distribution, achieving better colorization results.

3) TEXT-BASED COLORIZATION

These methods aim to leverage textual descriptions to associate objects in grayscale images with corresponding color words, generating colored images that align with user instructions. However, this approach often encounters challenges such as context confusion and spatial inconsistency. To address these issues and enhance the handling of complex sentences, Chen et al. [36] employs a recurrent attention model to spatially fuse image and language features, reinforcing the spatial consistency of colorization results. Xie [37] introduces an additional semantic segmentation task to facilitate the learning of higher-level semantics. Weng et al. [38], Chang et al. [39] relies on the correspondence between object words and color words in textual descriptions to guide the model in assigning specific color words to designated regions of grayscale images. Bahng et al. [40] proposes a method of generating a color palette based on text, which aids in managing the overall color distribution within images. In a recent study [13], an innovative image colorization framework is introduced. This model combines multiple conditions, including strokes, reference images, text, to direct the coloring of grayscale images. Despite these advancements, existing methods still exhibit certain drawbacks, such as the issue of colors in most images not appearing sufficiently natural and consistent.

III. PROPOSED METHODOLOGY

Drawing inspiration from the generation capabilities of the Stable Diffusion model, we aim to address the challenges in existing image colorization tasks, specifically balancing colorfulness and fidelity while preserving local spatial structures and eliminating the ghosting effect.

Our method leverages color-conditioned or text-conditioned guidance to capture generative diffusion prior, thereby producing diverse and accurate colored images in accordance with user instructions. To represent color conditions accurately and flexibly, we introduces the innovative concept of color hint points, as first proposed by UniColor [13]. These are annotated, color-labeled points on the target grayscale image that provide explicit color information for specific regions. Differing from the usual practice of solely feeding text inputs into stable diffusion models for image generation, our approach allows for multi-modal inputs of grayscale images with color hint points and optional texts containing color descriptions. This method facilitates the generation of the desired colorized output with greater precision and alignment with user preferences. Existing pre-trained T2I models have inherent limitations, especially in offering pixel-level control based on images. Additionally, balancing colorfulness and precision in the final output is challenging, and simply fine-tuning these models is insufficient to meet these goals.

To better leverage the advantages of Stable Diffusion for our image colorization task, we propose a novel method pipeline. As illustrated in Fig. 1, we can observe that, it consists of two main stages: the pre-processing stage and the colorization stage. The pre-processing stage is optional and is used when color hint points are added on the grayscale images by users before colorization. When there is no need for pre-processing, the process is equivalent to unconditional colorization. In the colorization stage, beyond the use of the pre-trained Stable Diffusion model, we incorporates two essential modules: an additional trainable pixel-level encoder and the controllable feature modulation block. The trainable pixel-level encoder is designed to capture detailed features from the input images, ensuring that the model has a rich representation of the image content. The controllable feature modulation block then integrates these extracted features of the grayscale image (with color hint points) with the corresponding output features, enabling precise and contextually appropriate colorization. By combining these modules, our method significantly enhances the capability of stable diffusion models to generate high-quality, diverse

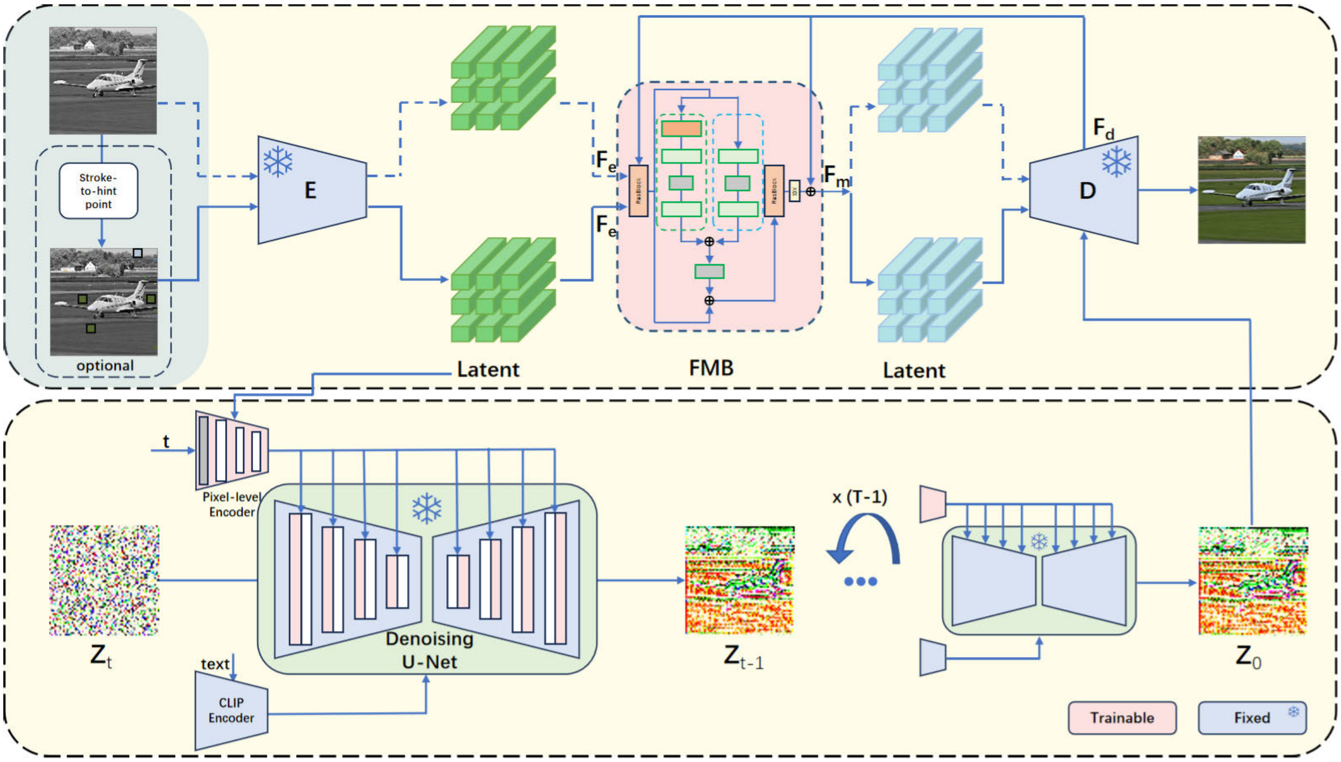


FIGURE 1. Framework overview. On pre-processing stage, We first adopt the Simple Linear Iterative Clustering (SLIC) algorithm [41] to get superpixel image and randomly choose several grid units to generate grayscale images with color hint points as the input of colorization stage. Then, on colorization stage, we introduce an additional trainable pixel-level encoder with time embedding to capture multi-scale features and temporal information, couple with text embedding from CLIP encoder, which guide the diffusion process to capture generative diffusion prior for image colorization. In order to strike a balance between colorfulness and precision, we propose a controllable feature modulation block consists of residual block and MS-CAM to modulate a fusion feature F_m by features from encoder F_e and features from decoder F_d.

colored images that closely adhere to user instructions, thus pushing the boundaries of what is achievable in image colorization tasks.

A. PRE-PROCESSING STAGE

During the pre-processing stage, we divide the image into a grid of $n \times n$. Each color hint point then corresponds to one of these single-color cells. The strokes drawn by users on the grayscale image guide our traversal through the cells. If the amount of colored pixels in a cell exceeds a certain threshold (e.g., 75% of the cell area), we associate the stroke's color with that color hint point. This process is iterated for all strokes, generating a grayscale image coupled with color hint points. This annotated image is subsequently used in the colorization stage, as illustrated in Fig. 1.

B. COLORIZATION STAGE

At this stage, we propose a framework that incorporates a pre-trained Stable Diffusion model as a color prior for image colorization. In crafting the model architecture, we address two essential considerations: 1) taking the grayscale image with color hint points from the pre-processing stage or textual descriptions as multi- modal input and the model is required to produce colored images that align with user instructions, ensuring both realism and vividness; 2) making minimal adjustments to the original Stable Diffusion model to maintain the integrity of its internal components. In order to meet these requirements, we introduce two supplementary principal modules in addition to the pre-trained Stable Diffusion model: a trainable pixel-level encoder and a controllable feature modulation block.

1) MULTI-MODAL INPUT

To enhance the vividness and realism of images generated by Stable Diffusion, we incorporate a time embedding layer into the trainable pixel-level encoder. This temporal information, along with the grayscale image (with color hint points) in the latent space, is fed into the encoder. The multi-scale features produced are then pass through a residual Block and into the denoising U-Net.

In conventional Stable Diffusion models, text features are typically obtained through a CLIP encoder and subsequently input into the denoising U-Net. We have improved upon this by modifying the SpatialTransformerV2, replacing the original Feed-Forward Network with a Gated-Dconv Feed-Forward Network. This enhancement improves the flexibility and accuracy of feature representation, thereby enhancing the quality and detail of the generated images.

2) PIXEL-LEVEL GUIDENCE

Due to the inherent randomness of the diffusion process, relying solely on a pre-trained Stable Diffusion model proves insufficient to achieve the task for image colorization. The challenge lies in enabling the diffusion process to perceive pixel-level image details and textures. It is well-known that models like ControlNet [8], T2I-Adapter [9], and others can effectively guide the diffusion process in achieving specific tasks based on different conditions (e.g., edges, sketches, canny), but they exhibit limitations in pixel-level control. For the purpose of guiding the generation process, We utilize an additional trainable encoder. The encoder has the same architecture as the encoder in the denoising U-Net of Stable Diffusion model and includes an additional time embedding layer. It extracts multi-scale features from the original grayscale image(with color hint points), denoted as F^n, which are subsequently transform to F_t^n using affine transformation layers.

$$\lambda^n, \mu^n = Conv_\theta^n \left(F^n\right) \tag{1}$$

$$\hat{F}_t^n = \mu^n + \left(\lambda^n + 1\right) \odot F_t^n \tag{2}$$

where $Conv_\theta^n (*)$represents a compact network composed of multiple convolutional layers and λ^n, μ^n represent the parameters of affine transformation layers.

During the training process, firstly, we fix the weight of the encoder and decoder from the pre-trained model. Additionally, we train the pixel-level encoder to extract multi-scale feature from the grayscale image (with color hint points) to guide the diffusion model in retaining diffusion prior.

3) CONTROLLABLE FEATURE MODULATION BLOCK

o strike a balance between colorfulness and precision. We have devised a trainable controllable feature modulation block, inspired by CodeFormer [42]. Let F_e and F_d represent the features from the encoder and decoder, respectively. Our objective is achieved by obtaining the transformed feature F_f in a residual manner. The fundamental concept involves harnessing encoder features from an autoencoder to modulate the corresponding decoder features. Here, we introduce an adjustable coefficient, denoted as $\alpha \in [0, 1]$, which serves to govern the degree of this adjustment, enabling a controlled feature modulation process.

$$F_m = \alpha \times M \left(F_e, F_d; \theta\right) + F_d \tag{3}$$

where $M (, ; \theta)$ denotes an improved Multi-scale Channel Attention Module with trainable parameter θ. Based on this design, we observe that a smaller α effectively exploits the generative capacity of stable diffusion, leading to output images with higher colorfulness. Conversely, a larger α relies more heavily on guidance from the grayscale input, thereby enhancing image fidelity. We find that a balanced compromise between colorfulness and precision is achieved when $\alpha = 0.75$.

IV. EXPERIMENTS

A. EXPERIMENTAL SETTING

1) DATASET

We conducted experiments on the extended COCO-Stuff dataset [38] and the Mini-ImageNet dataset [43]. The extended COCO-Stuff dataset derived from the original COCO-Stuff dataset [44]. This extended dataset comprises 59,000 training images and 2,400 evaluation images, curated by removing samples that did not meet the requirements for colorization tasks. Each image in the dataset is accompanied by a corresponding textual description that includes color information. The Mini-ImageNet dataset, which is a subset of the ImageNet dataset [45]. The Mini-ImageNet dataset consists of 100 randomly chosen categories from ImageNet1K. Each category includes 600 annotated images in the training set, amounting to a total of 60,000 images. The validation set contains 100 annotated images per category, resulting in a total of 10,000 images.

2) IMPLEMENTATION DETAILS

Pre-Processing Stage: During the pre-processing phase, drawing inspiration from unicolor [13], we adopt a method for synthesizing color hint points to generate grayscale images with color hint maps. Subsequently these images will serve as the input of the colorization stage. The fundamental concept of this approach lies in the random sampling of grid units from the original color image and the extraction of color hint points from each unit. For each chosen unit, the color hint points are sampled from $n \times n$ image grid cells. During the training phase, we start by employing the Simple Linear Iterative Clustering (SLIC) algorithm [41] to perform superpixel segmentation on the original color image I_c. Subsequently, by representing each superpixel with its average color, we obtain the superpixel image I_{sp}. Finally, we randomly choose several grid units from the superpixel image generated by SLIC algorithm, and the dominant color values within these units are utilized as the color hint points. Color reference regions are randomly selected from 30% to 50% of the quantized superpixel regions of the original image.During inference phase, users are allowed to draw strokes on the grayscale image. Then, employing SLIC algorithm to get the superpixel image. If the amount of colored pixels in a cell exceeds a predefined threshold (e.g., 75% of the cell area),, we associate the color of the stroke with that color hint point. To facilitate effective unconditional colorization during inference, there is a thirty percent probability of not introducing any color hint points during training or, with a fifty percent probability, randomly replacing the initial description with virtual captions such as “A Colorful Picture”, “A High-Quality Picture” or “ ”.

3) COLORIZATION STAGE

During the colorization phase, our framework is built upon the foundation of Stable Diffusion v2.1-base. We incorporate temporal information through the inclusion of a time embedding layer in the pixel-level encoder, which is the same as the encoder of the denoising U-Net in Stable Diffusion. Additionally, within each SpatialTransformerV2 block, we replace the original Feed-Forward Network with a Gated-Dconv Feed-Forward Network to achieve better control over the colorization process. We fine-tune the diffusion model over 20 epochs with a batch size of 8, using prompt that corresponds to the textual description of the images. We employ the Adam optimizer [46] with a learning rate set to 5×10^{-5}. Our training is conducted on images with a resolution of 512×512 using four A100 GPUs. For the controllable feature modulation block, we leverage the fine-tuned diffusion model to conditionally generate the corresponding latent codes z_0, given grayscale images with color hint points or textual descriptions. The training loss is closely aligned with the autoencoder used in Latent Diffusion Models (LDM) [11], with the primary distinction being the use of a fixed adversarial loss weight set to 0.025, instead of an adjustable weight.

4) EVALUATION METRICS

The fundamental evaluation factors for image colorization encompass perceptual realism and color vividness. To capture the essence of perceptual realism, we employ the Fréchet Inception Score (FID) [48] to measure the distribution similarity between predictions and ground truth. This metric provides valuable insights into how realistically the predictions align with the actual distribution. For evaluating color vividness, we turn to the Colorfulness metric [49], a measure akin to human vision perception. Besides, we also adopt the traditional metrics: PSNR [50], structural similarity index(SSIM) [51] and LPIPS [52], which are commonly used to assess the quality of generated images in comparison to the ground truth. It's crucial to acknowledge that plausible colorization outcomes might exhibit substantial color variations compared to the ground truth. Consequently, these metrics may not precisely depict the actual performance but should be regarded solely as reference points.

B. COMPARISON WITH STATE-OF-THE-ART

1) QUANTITATIVE COMPARISON

To demonstrate the effectiveness of our approach, we compared it against six state-of-the-art methods capable of coloring grayscale images without user prompts. These methods could be broadly categorized into the following three types: a) CNN-based methods: CIColor [4], Deoldify [39], InstColorization [3]; b) Transformer-based methods: ColorFormer [53], DDColor [54]; and c) GAN-based methods: BigColor [7]. For user-guided coloring, we compared our method with the latest and most representative transformer-based multimodal coloring method [13].We conducted comparisons of our method against these approaches on two datasets, namely the Mini-ImageNet [43] and the extended COCO-Stuff dataset [38], and quantified the results in Table 1. Although our method does not achieve the highest scores in PSNR, SSIM, and LPIPS metrics compared to DDColor, these metrics are not always the best indicators of colorization tasks, as they do not account for the diversity of colorization effects. In contrast, our method excels in terms of colorfulness, providing more vibrant and varied colorization results. Note that testing was performed using their official codebases and pre-trained weights.

2) QUALITATIVE COMPARISON

For automatic colorization, we show the image colorization results in Fig. 2. Apparently, as we can see that in the third-row images, the banana in the shopping cart is colored incorrectly by nearly all colorization methods showed, whereas our method accurately colors the banana in yellow. CIColor [4] and InstColorization [3] often suffer from color overflow or shortage issues, and latest methods like UniColor [13] and DDColor [54] may result in different parts of the same object being painted with different colors, disrupting the overall coherence of the image. Note that ground truth images are for reference only, although our method's coloring effect may not perfectly match the ground truth images, it can intelligently infer reasonable color schemes for most images, providing realistic colors. Additionally, it excels in edge handling, typically refining object boundaries to achieve natural color transitions with clear edges, thus reducing instances of color overflow or shortage. Our method maintains color consistency for the same objects in most cases, ensuring a stronger overall visual coherence and harmony. As for User-guided colorization, we divide it into two groups:color hint points guidance and text-based. For the former, we compare our method with Unicolor [13]. As depicted in Fig. 3, UniColor's generated horse lacks color richness as seen in the second-row image, and we can clearly observe that color bleeding occurs in the third-row image.In contrast, our method can effectively colors regions with color hint points accurately, and even for regions without color hint points, we can generate correct colorization results consistent with those guided by color hints. Moreover, Fig. 3 demonstrates that the colored images generated by our method are more natural and vivid, without issues of color bleeding. For the latter, we compare our method with L-CAD [47] and Unicolor [13]. As depicted in Fig. 4,

C. USER STUDY

In order to further validate our model's effectiveness, we conducted a user study, with the experimental findings detailed in this section. For unconditional colorization, we selected InstColorization [3] from CNN-based methods and ColorFormer [53] from transformer-based methods. As for user-guided colorization, we chose UniColor [13] and L-CAD [47]. In each scenario, we selected 25 images from the extended COCO-Stuff validation dataset [38] randomly. Participants were presented with a series of colorized images generated by these five methods and their corresponding Ground Truth (in the case of unconditional coloring). The images were then shuffled randomly and presented to

TABLE 1. Quantitative results with SOTA methods on benchmark datasets. ↑(↓) means higher (lower) is better.

	COCO-Stuff [38]					Mini-ImageNet [43]				
Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	CF↑	PSNR↑	SSIM↑	LPIPS↓	FID↓	CF↑
CIColor	22.91	0.938	0.211	20.41	34.28	21.79	0.946	0.205	14.26	33.87
Deoldify	23.22	0.941	0.178	11.06	26.53	23.65	0.950	0.173	8.57	22.20
InstColorization	23.74	0.949	0.163	10.06	27.36	22.56	0.919	0.185	14.33	24.76
BigColor	20.80	0.931	0.202	8.86	39.64	20.95	0.940	0.198	7.10	39.19
UniColor	22.37	0.940	0.185	7.85	33.32	21.69	0.948	0.202	8.62	32.89
ColorFormer	22.53	0.984	0.183	8.91	37.03	**23.08**	0.984	0.183	6.34	37.37
DDColor	23.04	**0.992**	**0.147**	**6.39**	40.63	22.63	**0.991**	**0.156**	**5.41**	41.46
Ours	**24.11**	0.974	0.167	7.33	**41.45**	22.54	0.943	0.183	7.34	**41.74**

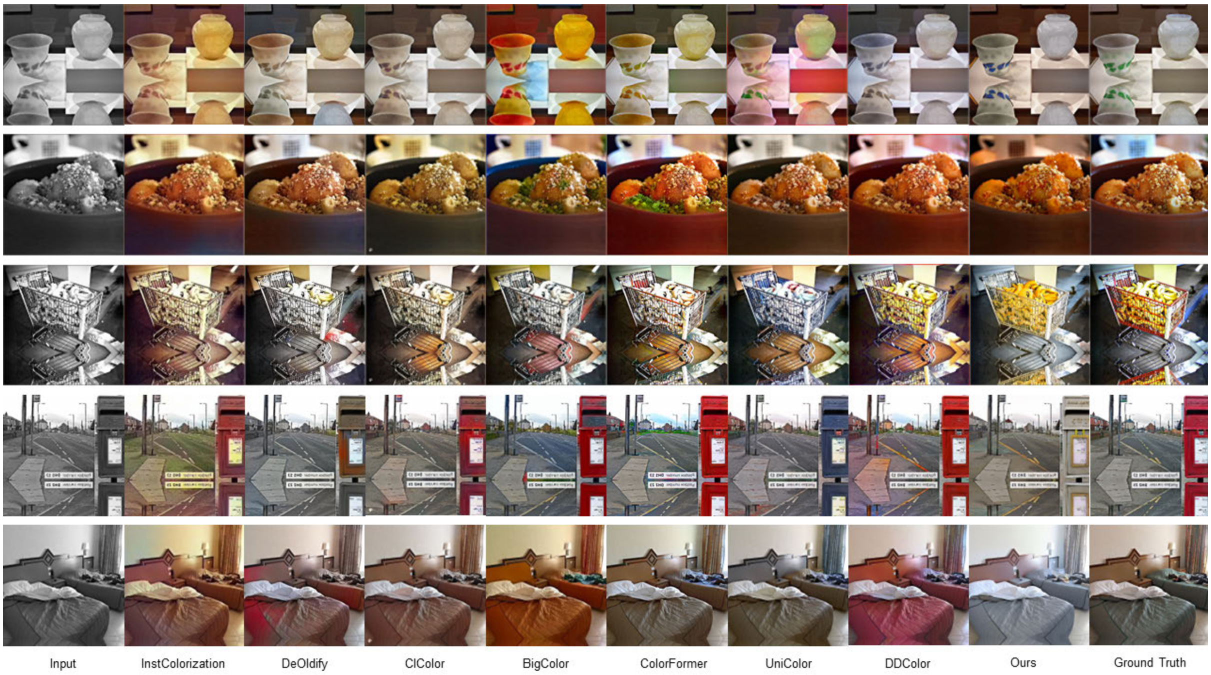


FIGURE 2. Visual comparison with seven existing image colorization methods.

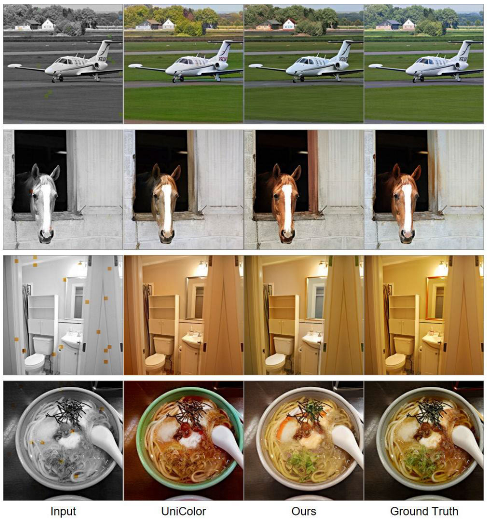


FIGURE 3. Visual comparison with the hint point-based method. (UniColor [13]).

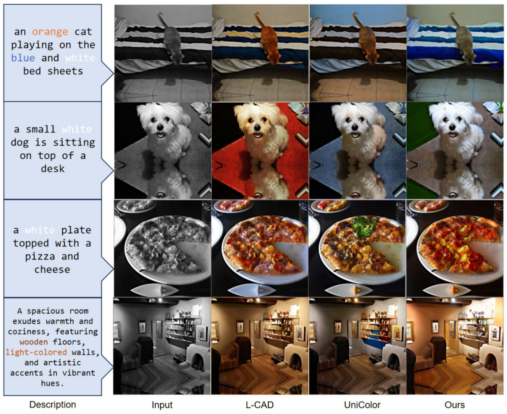

an orange cat playing on the blue and white bed sheets
a small white dog is sitting on top of a desk
a white plate topped with a pizza and cheese
A spacious room exudes warmth and coziness, featuring wooden floors, light-colored walls, and artistic accents in vibrant hues.
Description
Input
L-CAD
UniColor
Ours

FIGURE 4. Visual comparison with the text-based method.(UniColor [13] and L-CAD [47]).

participants, who were asked to pick the most realistic colorization and the one that most consistent with the input prompts (if applicable). 100 participants completed the evaluation. The results are shown in Fig. 5. As can be seen in the figure, in terms of unconditional colorization, all methods still fall short of reaching the level of real images, whereas

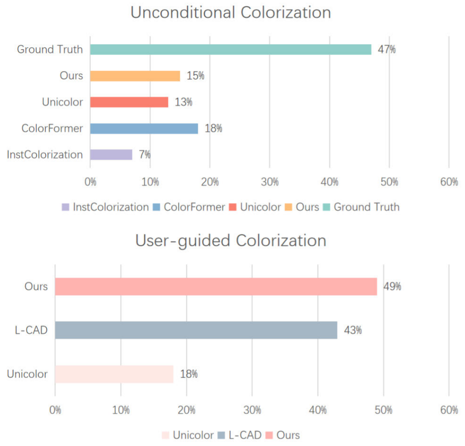


FIGURE 5. Results of user study on seven existing image colorization methods.

TABLE 2. Sensitivity Analysis of Coefficient α.

α	PSNR↑	SSIM↑	LPIPS↓	FID↓	CF↑
$\alpha = 0.0$	23.21	0.983	0.186	7.41	40.77
$\alpha = 0.5$	24.03	0.962	0.175	7.59	40.28
$\alpha = 0.75$	**24.11**	0.974	0.167	**7.33**	**41.45**
$\alpha = 1.0$	23.46	**0.955**	**0.163**	7.84	40.13

in the user-guided colorization scenario (using color hint points), our method significantly outperforms UniColor [13].

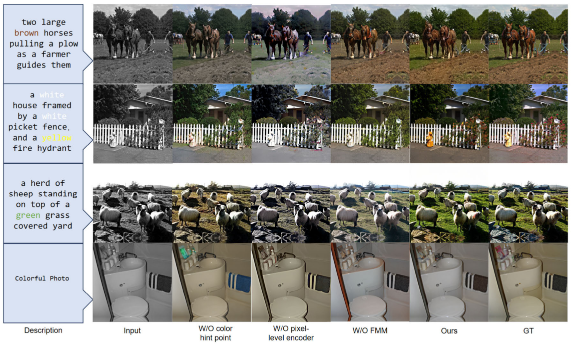

two large brown horses pulling a plow as a farmer guides them
a white house framed by a white picket fence, and a yellow fire hydrant
a herd of sheep standing on top of a green grass covered yard
Colorful Photo
Description
Input
W/O color hint point
W/O pixel-level encoder
W/O FMM
Ours
GT

FIGURE 6. Visual comparison with the text-driven editing methods.

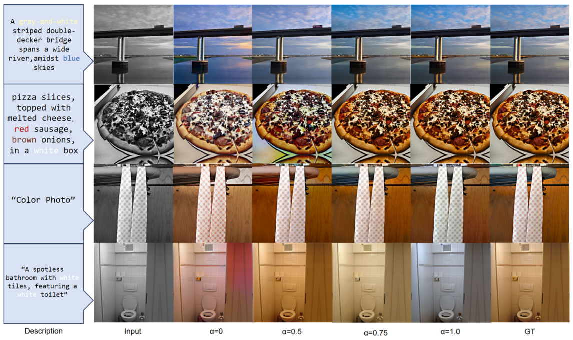

A gray-and-white striped double-decker bridge spans a wide river,amidst blue skies
pizza slices, topped with melted cheese, red sausage, brown onions, in a white box
"Color Photo"
"A spotless bathroom with white tiles, featuring a white toilet"
Description
Input
α=0
α=0.5
α=0.75
α=1.0
GT

FIGURE 7. Visual comparison with different coefficient α.

D. ABLATION STUDY

1) COLOR HINT POINTS

We input both the original grayscale image and the grayscale image with color hint points obtained after pre-processing into our colorization pipeline. As shown in Fig. 6, it is evident that the grayscale image with color hint points obtained after pre-processing achieves superior colorization performance, compared to the original grayscale image.

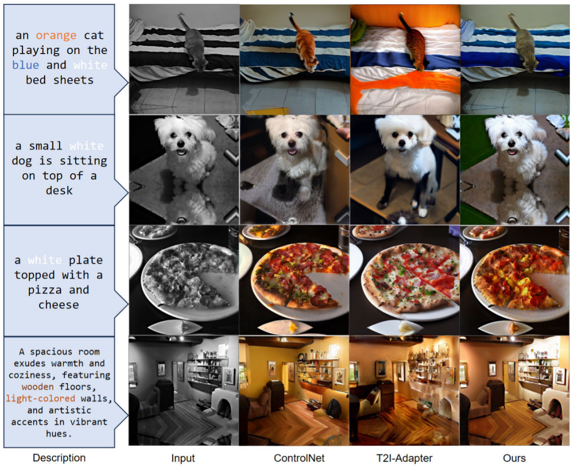

an orange cat playing on the blue and white bed sheets
a small white dog is sitting on top of a desk
a white plate topped with a pizza and cheese
A spacious room exudes warmth and coziness, featuring wooden floors, light-colored walls, and artistic accents in vibrant hues.
Description
Input
ControlNet
T2I-Adapter
Ours

FIGURE 8. Visual comparison with the text-driven editing methods.

2) TRAINABLE PIXEL-LEVEL ENCODER

Firstly, We investigate the importance of the trainable pixel-level encoder along with time embedding. Fig. 6 indicates that without the use of the encoder, the diffusion model can hardly maintain spatial structures of the original grayscale image due to lacking of the guidance of multi-scale features.

3) CONTROLLABLE FEATURE MODULATION BLOCK

Inspired by CodeFormer, our trainable controllable feature modulation block operates with a controllable coefficient $\alpha \in [0,1]$. A smaller α maximizes the generative capacity of stable diffusion, resulting in output images with higher colorfulness. Conversely, a larger α relies more on guidance from the grayscale input, thereby enhancing image precision. We present the qualitative results in Fig. 7 and the quantitative results in Table 2. In Fig. 7, We can observe that models with α (e.g., 0.75, 0.5) present better colorization outputs. As shown in Table 2, compared to $\alpha = 0.0$, models with larger α (e.g., 0.75) achieve lower FID scores which indicates better fidelity. On the other hand, $\alpha = 0.0$ achieves higher Colorfulness scores, suggesting greater vividness. Thus, an appropriate α can strike a balance between colorfulness and precision.

4) DISCUSSION

It is well-known that Recent studies like ControlNet [8], T2I-Adapter [9], and others can effectively guide the pre-trained diffusion model in achieving specific tasks based on different conditions (e.g., edges, sketches, text), but they exhibit limitations in pixel-level control. Because they are not specifically designed for the colorization task, leading to difficulties in maintaining local spatial structures. Consequently, these limitations hinder their ability to produce desired colorized outputs in accordance to users' will. We share our comparison in Fig. 8

V. CONCLUSION

We introduce a novel image colorization pipeline based on the Stable Diffusion v2.1-base model that effectively addresses the key challenges of balancing color saturation, visual coherence, and object-level semantics while minimizing artifacts. Our dual-stage pipeline, including pre-processing and colorization stages. We first generate grayscale images with color hint points by SLIC algorithm from user-drawn strokes in the pre-processing stage, while in the colorization stage, we utilize a trainable pixel-level encoder with time embedding to extract multi-scale features and temporal information, along with a controllable feature modulation block to achieve a balance between colorfulness and precision. All of these advancements underscore the potential of diffusion models to produce highly realistic and contextually accurate colorized images according to user instructions, offering significant contributions to image colorization.

Future Work: As we know, a major barrier to the practical adoption of diffusion models is their sampling speed. Our diffusion pipeline also requires additional sampling processes to visualize the coloring results, which significantly slows down the generation speed and limits the potential for real-time applications. To address this issue, future work may focus on integrating progressive distillation methods. This approach has the potential to drastically reduce sampling times while maintaining high-quality results, thereby making real-time applications more feasible.

REFERENCES

[1] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 415–423.

[2] A. Deshpande, J. Lu, M.-C. Yeh, M. J. Chong, and D. Forsyth, "Learning diverse image colorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2877–2885.

[3] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware image colorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7965–7974.

[4] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 649–666.

[5] J. Antic. (2019). *Jantic/Deoldify: A Deep Learning Based Project for Colorizing and Restoring Old Images (and Video!)*. [Online]. Available: https://github.com/jantic/DeOldify

[6] Y. Wu, X. Wang, Y. Li, H. Zhang, X. Zhao, and Y. Shan, "Towards vivid and diverse image colorization with generative color prior," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14357–14366.

[7] 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 *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 350–366.

[8] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3836–3847.

[9] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2I-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," 2023, *arXiv:2302.08453*.

[10] A. Voynov, K. Aberman, and D. Cohen-Or, "Sketch-guided text-to-image diffusion models," in *Proc. Special Interest Group Comput. Graph. Interact. Techn. Conf. Conf.*, Jul. 2023, pp. 1–11.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.

[12] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.

[13] Z. Huang, N. Zhao, and J. Liao, "UniColor: A unified framework for multi-modal colorization with transformer," *ACM Trans. Graph.*, vol. 41, no. 6, pp. 1–16, Dec. 2022.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, p. 4.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[17] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[18] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6877–6886.

[19] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2021, pp. 17864–17875.

[20] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," 2021, *arXiv:2102.04432*.

[21] S. Weng, J. Sun, Y. Li, S. Li, and B. Shi, "CT^2: Colorization transformer via color tokens," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 1–16.

[22] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," in *Proc. ACM SIGGRAPH Papers*, Aug. 2004, pp. 689–694.

[23] Y. Endo, S. Iizuka, Y. Kanamori, and J. Mitani, "DeepProp: Extracting deep features from a single image for edit propagation," *Comput. Graph. Forum*, vol. 35, no. 2, pp. 189–201, May 2016.

[24] Y. Xiao, P. Zhou, Y. Zheng, and C.-S. Leung, "Interactive deep colorization using simultaneous global and local inputs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1887–1891.

[25] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros, "Real-time user-guided image colorization with learned deep priors," 2017, *arXiv:1705.02999*.

[26] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," in *Proc. 29th Annu. Conf. Comput. Graph. Interact. Techn.*, Jul. 2002, pp. 277–280.

[27] X. Liu, L. Wan, Y. Qu, T.-T. Wong, S. Lin, C.-S. Leung, and P.-A. Heng, "Intrinsic colorization," in *Proc. ACM SIGGRAPH Asia Papers*, Dec. 2008, pp. 1–9.

[28] A. Bugeau, V.-T. Ta, and N. Papadakis, "Variational exemplar-based image colorization," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 298–307, Jan. 2014.

[29] Y.-W. Tai, J. Jia, and C.-K. Tang, "Local color transfer via probabilistic segmentation by expectation-maximization," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2005, pp. 747–754.

[30] R. Ironi, D. Cohen-Or, and D. Lischinski, "Colorization by example," *Rendering Techn.*, vol. 29, pp. 201–210, Jun. 2005.

[31] A. Y.-S. Chia, S. Zhuo, R. K. Gupta, Y.-W. Tai, S.-Y. Cho, P. Tan, and S. Lin, "Semantic colorization with Internet images," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 1–8, Dec. 2011.

[32] R. K. Gupta, A. Y.-S. Chia, D. Rajan, E. S. Ng, and H. Zhiyong, "Image colorization using similar images," in *Proc. 20th ACM Int. Conf. Multimedia*, Oct. 2012, pp. 369–378.

[33] X. Dong, W. Li, X. Wang, and Y. Wang, "Learning a deep convolutional network for colorization in monochrome-color dual-lens system," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8255–8262.

[34] X. Dong, W. Li, X. Wang, and Y. Wang, "Cycle-CNN for colorization towards real monochrome-color camera systems," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 10721–10728.

[35] P. Lu, J. Yu, X. Peng, Z. Zhao, and X. Wang, "Gray2ColorNet: Transfer more colors from reference image," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3210–3218.

[36] J. Chen, Y. Shen, J. Gao, J. Liu, and X. Liu, "Language-based image editing with recurrent attentive models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8721–8729.

[37] Y. Xie, "Language-guided image colorization," M.S. thesis, Dept. Comput. Sci., ETH Zurich, Zurich, Switzerland, 2018.

[38] S. Weng, H. Wu, Z. Chang, J. Tang, S. Li, and B. Shi, "L-CoDe: Language-based colorization using color-object decoupled conditions," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 3, pp. 2677–2684.

[39] Z. Chang, S. Weng, Y. Li, S. Li, and B. Shi, "L-CoDer: Language-based colorization with color-object decoupling transformer," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 360–375.

[40] H. Bahng, S. Yoo, W. Cho, D. K. Park, Z. Wu, X. Ma, and J. Choo, "Coloring with words: Guiding image colorization through text-based palette generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 431–447.

[41] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[42] G. Liu, X. Zhou, J. Pang, F. Yue, W. Liu, and J. Wang, "Codeformer: A GNN-nested transformer model for binary code similarity detection," *Electronics*, vol. 12, no. 7, p. 1722, Apr. 2023.

[43] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–11.

[44] H. Caesar, J. Uijlings, and V. Ferrari, "COCO-stuff: Thing and stuff classes in context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1209–1218.

[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[47] 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 *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 3–6.

[48] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, p. 6.

[49] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Human Vision and Electronic Imaging VIII*, vol. 5007. Bellingham, WA, USA: SPIE, 2003, pp. 87–95.

[50] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electron. Lett.*, vol. 44, no. 13, pp. 800–801, 2008.

[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

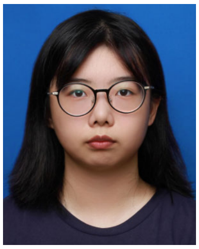
[52] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[53] X. Ji, B. Jiang, D. Luo, G. Tao, W. Chu, Z. Xie, C. Wang, and Y. Tai, "ColorFormer: Image colorization via color memory assisted hybrid-attention transformer," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 20–36.

[54] X. Kang, T. Yang, W. Ouyang, P. Ren, L. Li, and X. Xie, "DDColor: Towards photo-realistic and semantic-aware image colorization via dual decoders," *arXiv preprint arXiv:2212.11613*, 2022.

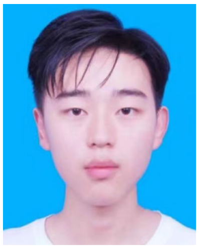
WANYU YANG received the B.S.E. degree in mechanical design, manufacturing and automation from Shanghai Normal University. She is currently pursuing the M.S. degree in digital media creative engineering with Shanghai Film Academy, Shanghai University, China. Her research interests include image generation and leveraging deep learning methodologies.

FEIFAN CAI received the B.S. degree in science from Changsha University of Technology, in 2019, and the M.S. degree in communication and information engineering from Shanghai University, China, in 2022, where he is currently pursuing the Ph.D. degree in digital media creative engineering with Shanghai Film Academy. His research interests include image and video super-resolution reconstruction, and leveraging deep learning methodologies.



YANG SHU received the B.S.E. degree from Anhui University of Science and Technology, in 2022. She is currently pursuing the M.S. degree in computer science and technology with Shanghai Film Academy, Shanghai University, China. Her research interests include image restoration, super-resolution reconstruction, and leveraging deep learning methodologies.

ZIHAO ZHANG received the bachelor's degree from Foshan University. He is currently pursuing the master's degree with Shanghai University. His current research interests include low-level vision and deep learning.



QI LIU received the B.S. degree from Nanjing University of Posts and Telecommunications, in 2023. He is currently pursuing the master's degree in digital media creative engineering with Shanghai Film Academy, Shanghai University, China. His research interests include image and animation super-resolution reconstruction, and leveraging deep learning methodologies.

YOUDONG DING received the Ph.D. degree in mathematics from the University of Science and Technology of China, Hefei, China, in 1997. From 1997 to 1999, he was a Postdoctoral Researcher with the Department of Mathematics, Fudan University, Shanghai, China. He is currently a Professor with Shanghai University, Shanghai. His research interests include computer graphics, image processing, and digital media technology.