COCO-LC: Colorfulness Controllable Language-based Colorization

Yifan Li
Wangxuan Institute of Computer Technology,
Peking Universiy, Beijing, China
2100012520@stu.pku.edu.cn

Yuhang Bai
Wangxuan Institute of Computer Technology,
Peking Universiy, Beijing, China
20194779@stu.neu.edu.cn

Shuai Yang
Wangxuan Institute of Computer Technology,
Peking Universiy, Beijing, China
williamyang@pku.edu.cn

Jiaying Liu*
Wangxuan Institute of Computer Technology,
Peking Universiy, Beijing, China
liujiaying@pku.edu.cn

(a) our language-based image colorization results

(b) controllable colorization

Figure 1: We propose COCO-LC, a novel colorfulness controllable language-based colorization framework. (a) COCO-LC generates realistic and semantic-consistent colorization results. (b) COCO-LC allows for flexible user control over (top) color types and (bottom) color styles.

Abstract

Language-based image colorization aims to convert grayscale images to plausible and visually pleasing color images with language guidance, enjoying wide applications in historical photo restoration and the film industry. Existing methods mainly leverage large language models and diffusion models to incorporate language guidance into the colorization process. However, it is still a great challenge to build accurate correspondence between the gray image and the semantic instructions, leading to mismatched, overflowing and under-saturated colors. In this paper, we introduce a novel coarse-to-fine framework, COlorfulness COntrollable Language-based Colorization (COCO-LC), that effectively reinforces the image-text correspondence with coarsely colorized results. In addition, a multi-level condition that leverages both low-level and high-level cues of the gray image is introduced to realize accurate semantic-aware colorization without overflowing colors. Furthermore, we condition COCO-LC with a scale factor to determine the colorfulness of the output, flexibly meeting the different needs of users. We validate the superiority of COCO-LC over state-of-the-art image colorization methods in accurate, realistic and controllable colorization through extensive experiments. The code and demo are available at *https://lyf1212.github.io/COCO-LC*.

CCS Concepts

• **Computing methodologies** → **Reconstruction**; **Image processing**; **Computational photography**.

Keywords

Colorfulness control, language guidance, image colorization

ACM Reference Format:
Yifan Li, Yuhang Bai, Shuai Yang, and Jiaying Liu. 2024. COCO-LC: Colorfulness Controllable Language-based Colorization. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3680662

1 Introduction

Color plays a pivotal role in shaping human perception of the world. It serves not only as one of the most expressive visual elements, but also directly influences people's emotions through various color styles. Image colorization aims to convert grayscale images into

*Corresponding author

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3680662

color images, which has a wide applications in diverse fields such as old photo restoration, color grading, and automatic animation colorization. However, image colorization is an ill-posed problem as it involves inferring a three-channel color image from a single-channel illuminance image, which may have multiple reasonable solutions. Despite many automatic colorization methods [9, 20, 23, 25, 36, 49] have been proposed, most of them suffer such color ambiguity, leading to under-saturared colors with the mean of all possible color choices. To address this issue, conditional colorization begins to attract researchers' interests. By imposing additional constraints, such as language [5–7, 17, 42, 47], scribble [17, 46], reference image [27] and palette [39, 43], the method can more accurately render specified colors.

Recent advancement on large language models (LLMs) and diffusion models empowers image processing with language guidance. Compared with other visual conditions, text descriptions are highly informative, and are simple and efficient to use. Recent language-based image colorization methods [5, 7, 17, 42] either train language-vision aligned models supervised by dense human annotations, which may overfit on a small dataset and result in poor quality; or employ pretrained large models such as BERT [10] and CLIP [31] to align language domain and image domain, which, however, is less effective in building an accurate correspondence between gray instances and text description.

Based on the above analysis, we summarize three key requirements lies on the image colorization problem: **(1) Realism.** The colorization results should look realistic. **(2) Consistency.** The colorization results should match the semantic content of the original gray images. **(3) Controllability.** Users can flexibly adjust the color of the output. Existing methods can only fulfill one or two requirements, hardly satisfying all above standards: The unconditional ones provide no controllability and suffer under-saturation, while the conditional ones are less flexible or fail to render semantically correct colors, leading to color overflow and inconsistency.

To build a powerful image colorization framework that simulatenously achieves realism, consistency and controllability, we propose **COCO-LC** for **CO**lorfulness **CO**ntrollable **L**anguage-based **C**olorization. To strengthen realism, our key idea is a coarse-to-fine framework that leverages Stable Diffusion [34], the pre-trained cross-modality generative model, to utilize its high capability for textual-visual modeling and powerful generative prior on the natural image distribution. To maintain the semantic consistency between gray inputs and colorization results, we further incorporate robust multi-level conditions with both low-level and high-level cues of the gray inputs. Finally, in order to provide more controllability, besides color types, we develop a novel colorfulness-controllable colorization decoder to produce results with diverse color styles ranging from vintage, realistic, to fantastic.

Specifically, we propose a coarse-to-fine training framework. In the coarse colorization stage, we inject features from the large-scale cross-modality model CLIP [31] into the latent space of the gray image. On this basis, the resulting latent codes with rich color information are used to guide the diffusion model to generate fine-level colorization results, which achieves accurate correspondence between the textual color words and visual gray instances.

Additionally, we propose multi-level conditions that effectively alleviate color overflow and mismatch issues. We design a novel dual-branch feature extractor that aligns the feature granularity of the low-level edge features with diffusion features for balanced condition injection. A semantic-aware feature regularization is further proposed to provide high-level features to improve semantical correspondence.

In terms of controllability, brightness, hue and saturation are three key characteristics to determine color values. While brightness is determined by the gray image and the control of hue is well studied for conditional colorization, controlling saturation remains less explored. To this end, we design a colorfulness-controllable colorization decoder, with a scaling factor to allow users to flexibly choose different color styles from vintage to gorgeous styles.

With the novel designs above, COCO-LC comprehensively achieves realism, consistency, and controllability. Extensive experimental results demonstrate our superiority in generating high-quality color images over both automatic and language-based state-of-the-art baselines. In summary, our contributions are threefold:

- We propose a novel coarse-to-fine COCO-LC framework for language-based colorization that achieves high realism, consistency, and controllability simultaneously. We leverage cross-modality model CLIP to build accurate correspondence between the color words and gray instances, significantly improving the visual-textual color consistency.
- We develop multi-level conditions with both low-level and high-level guidance to find accurate color-semantic correspondence. A novel dual-branch feature extractor is proposed to align the feature granularity for balanced condition injection, which effectively alleviates color overflow.
- We design a colorfulness-controllable decoder, which adaptivelty fuses the predicted color information and the original grayscale information, allowing users to choose fantastic, realistic or vintage colorized results to flexibly satisfy diverse user preferences.

2 Related works

2.1 Automatic Colorization

Automatic colorization aims to generate plausible colorful images from grayscale input without user guidance. Cheng *et al.* [9] introduced the first deep-based colorization model. In view of the imbalance of the color space, CIC [49] defines a classification loss in the quantized *ab* color space rather than the traditional regression loss, leading to more saturated results. Researchers produce models with more plausible colorization results with Pixcolor [12]. With the development of generative models, many researchers began to make efforts on colorization problem with GAN [11]. Cao *et al.* [4] conditioned GAN with grayscale information in multiple layers to maintain spatial feature consistency. ChromaGAN [38] adds color error and class distribution losses to optimize training for colorization. BigColor [23] improves vividness through pre-trained generative priors while suffering from color cast. DeOldify [1] utilizes U-Net as the generator model with a self-attention scheme and reduces direct GAN training time with a scheduled adversarial loss. Furthermore, to utilize the long range dependency and increase model capacity, transformer-based architectures [37] are widely used in colorization tasks [20, 21, 25]. ColorFormer [20] designs a color memory module to learn and store the semantic-color

mapping. HistoryNet [21] introduces fine-grained semantic understanding and classification prior to achieve accurate colorization and prevent color overflow. Priors including class labels [19], instance bounding boxes [36], and semantic segmentation maps [51] are introduced to further guide colorization models with semantic information. However, unconditional colorization is an ill-posed problem, resulting in color ambiguity and under-saturation. From the users' perspective, lack of controllability limits the practicability of this kind of approaches.

2.2 Language-based Colorization

With the help of the flexibility of text descriptions, language-based colorization methods enable simple but effective control over instance colors. Unicolor [17] developed a unified framework to support colorization in multiple modalities, including text descriptions. It proposed a spatial partitioning heuristic method to align color words and instances by leveraging CLIP [31] as zero-shot classifier, but struggle with finer colorization ability in spatial, and can only deal with several classic colors. L-CoDe [42] and L-CoDer [5] both utilized an color-object correspondence matrix to decouple color and instance, along with a soft-gated injection module for resolving the color-object mismatch issue. However, both L-CoDe and L-CoDer need to create an additional color-instance decoupling module to align colors with gray instances. This module have to train on extra human annotations, resulting in more training cost. Thus, L-CoIns [7] achieved instance awareness with the grouping mechanism to adaptively aggregate similar image patches without additional annotations. Furthermore, L-CAD [6] reduces the dependence of colorization on the precision of language guidance by using a pretrained diffusion model with rich color priors and superior cross-modal capability. However, it lacks the ability to generate high-resolution colorized images due to the cost of fine-tuning.

2.3 Diffusion-based Image Colorization

Recently, diffusion model shows its advantages in image generation [32, 34], image editing [35] and image restoration [29, 45]. As diffusion models hold high-quality color and semantic priors with the benifit from large-scale pretraining, L-CAD [6] designs a luminance-guided compression module and merge latent codes of grayscale images into diffusion process to maintain accurate structure when fine-tuning the Stable Diffusion [34]. Diffusing Colors [47] adopts the cold diffusion mechanism [2] to learn a colorization process iteratively. Besides, CtrlColor [28] develops multiple encoders based on ControlNet [48] according to different kind of conditions, such as user scribbles, reference images, regions or text descriptions. As colorized results need to spatially align with input grayscale images, diffusion-based colorization methods have to balance the trade-off between realism and consistency. Overall, diffusion-based image colorization is still a cutting-edge and challenging topic.

3 COCO-LC

3.1 Preliminary: Stable Diffusion

Diffusion models learn image distributions based on a diffusion process and a denoising process, where Stable Diffusion [34] operates in the latent space with a VAE encoder $\mathcal{E}$ and a VAE decoder $\mathcal{D}$. During the diffusion process, random Gaussian noises $\epsilon \sim \mathcal{N}(0, I)$ are gradually added to the encoded latent feature $z_0 = \mathcal{E}(x_0)$ of the input image x_0 in T steps,

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t}\epsilon, \tag{1}$$

producing a series of noisy samples $z_1, ..., z_T$. As z_T can be treated as a standard Gaussian noise approximately when T is large enough, the denoising process can recover a realistic image $x_0 = \mathcal{D}(z_0)$ from a srandard Gaussian noise iteratively, which is achieved by training a neural network U-Net ϵ_θ with parameter θ to predict ϵ at each timestep t based on z_t with the loss function:

$$\mathcal{L} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I),t}[\|\epsilon - \epsilon_\theta(z_t, c_{text}, t)\|^2], \tag{2}$$

where in Stable Diffusion, text conditions c_{text} are taken as guidance to constrain the generation through cross-attention mechanism. Besides text, Stable Diffusion can be addtionally conditioned on images with ControlNet [48]. ControlNet is a trainable copy of the diffusion model, serving as a side branch to accept and apply image conditions c_i to the main diffusion branch. The overall learning objective can be formulated as

$$\mathcal{L} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I),t}[\|\epsilon - \epsilon_\theta(z_t, c_{text}, c_i, t)\|^2]. \tag{3}$$

It is natural to apply ControlNet to the colorization task, *i.e.*, using grayscale image I_{gray} as c_i to predict the corresponding colorful image I. However, we experimentally found that vanilla ControlNet is not competent to offer high-quality gray image constraints. ControlNet uses simple convolution layers to preprocess the conditional images, leading to imbalanced feature granularity between the ControlNet and diffusion branches and causing structure distortion. It also fails to build robust correspondence between the gray instances and the color words, resulting in color mismatch and color overflow. This paper propose a new croase-to-fine framework with novel multi-level consistency-aware conditions to solve the above problems.

3.2 Overview architecture

COCO-LC takes as input a grayscale image I_{gray}, a text prompt describing the desired color of the instances in c_{text}, and a scaling factor α to indicate the targeted colorfulness, and produce a corresponding colorization result $\hat{I}$. As shown in Fig. 2, the proposed COCO-LC framework consist of three key components: (1) coarse-to-fine colorization; (2) consistency-aware multi-level condition; (3) colorfulness controllable colorization decoder.

Our coarse-to-fine colorization (Sec. 3.3) first colorizes the latent feature $z_{gray} = \mathcal{E}(I_{gray})$ with our proposed lightweight Colorful Information Adaptor (CIA). CIA leverages the power of CLIP [31] to create a coarsely colorized feature based on c_{text} and z_{gray}. The resulting $\hat{z}$ is fed into the main diffusion branch by our condition injection branch for fine-level colorization.

For consistency-aware multi-level condition, we extract the edge map and segmentation map from I_{gray} to serve as low-level structure and high-level semantic guidance, respectively. We merge the low-level edge maps and high-level semantic segmentation maps with the condition injection branch, as will be detailed in Sec. 3.4.

In Sec. 3.5, we will introduce our colorfulness controllable colorization decoder (COCO-decoder), that adapts the VAE decoder $\mathcal{D}$ to merge grayscale image and provides a scaling factor α to enable users to flexibly adjust the colorfulness of the output.

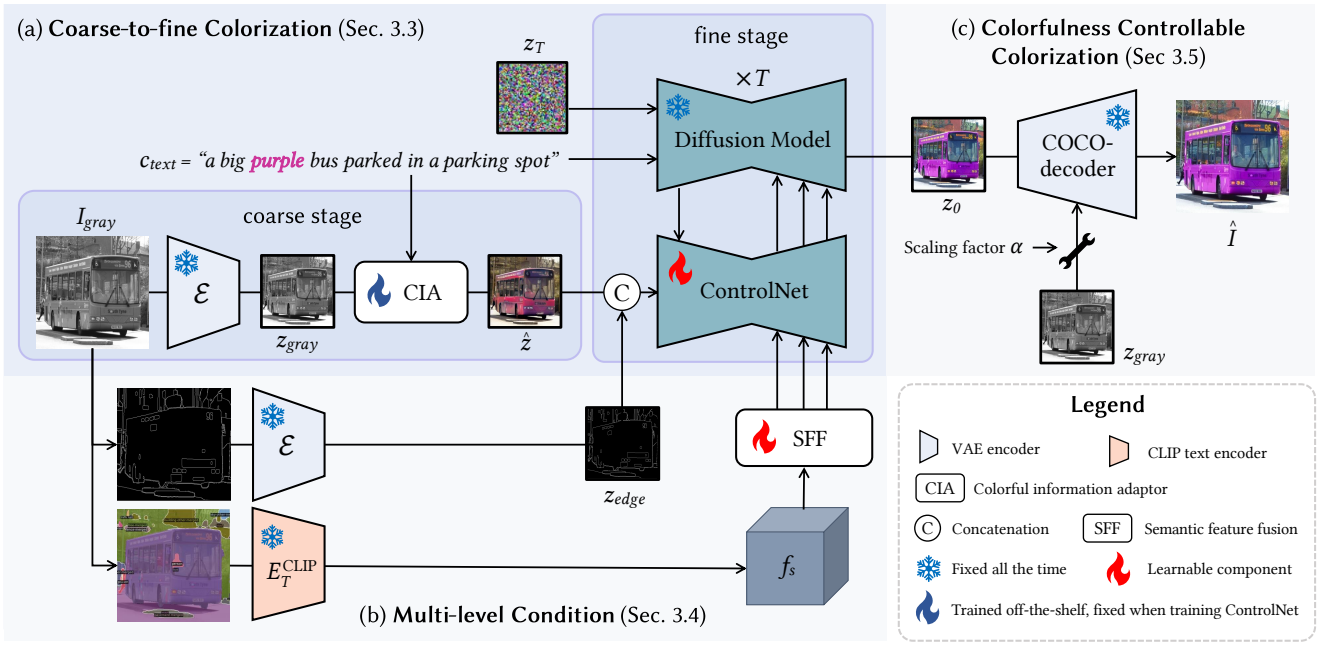

Figure 2: Illustration of the proposed COCO-LC framework with three key components: (a) coarse-to-fine colorization; (b) consistency-aware multi-level condition; (c) colorfulness controllable colorization.

3.3 Coarse-to-fine Colorization Framework

While Stable Diffusion demonstrates satisfying performance of matching color words and image instances with cross-attention mechanism, it is not trival to find proper correspondence when the instances lie in the gray feature space because of the domain gap between the grayscale images and the color images. Our coarse-to-fine colorization framework solve this issue by gradually narrow the domain gap in two stages.

3.3.1 Coarse colorization. We design a Colorful Information Adaptor (CIA) to merge color information into VAE latent space, as shown in Fig. 3. This lightweight adaptor is trained off-the-shelf and kept fixed during fine colorization stage, enables its flexibility and simplicity. During training, given a color image I, we obtain its grayscale version I_{gray} and map it into feature domains with VAE encoder $\mathcal{E}$. For its corresponding text prompt $\mathcal{P}$, we utilize CLIP text encoder E_T^{CLIP} to get text embedding $f_T = E_T^{CLIP}(\mathcal{P})$. Taking gray latents $\mathcal{E}(I_{gray})$ and CLIP color text embeddings f_T as input, CIA merges the grayscale contents and color hints through a scale-shift operation:

$$\hat{z} = \text{GN}(z) \cdot \big(1 + F_{\text{scale}}(f_T)\big) + F_{\text{shift}}(f_T), \tag{4}$$

where $F_{\text{scale}}, F_{\text{shift}}$ denote the mapping networks in CIA to wrap the color information into scale-shift parameters. z is the structure feature initialized with $\mathcal{E}(I_{gray})$. GN denotes group normalization [44]. This fusion operation can be also understood as a parameterized version of AdaIN [16], and is applied to z twelve times with six sequenced ResBlocks [14].

Except for basic MSE loss and perceptual loss $\mathcal{L}_{\text{per}}$, we further develop a regularization on chrominance channels for more saturated coarse results. Drawing inspiration from [41, 49] which construct empirical probability on quantized ab space, we first downsample coarse results $I_{\text{coarse}} = \mathcal{D}(\text{CIA}(\mathcal{E}(I_{gray}), f_T))$ to the resolution of 64, assigning color categories to each pixels and employing a cross-entropy loss on quantized color classes:

$$\mathcal{L}_{\text{ab}} = \text{CE}\big(\boldsymbol{Q}(I_{\text{coarse}}), \boldsymbol{Q}(I)\big), \tag{5}$$

where $\boldsymbol{Q}$ denotes the operation of quantization, CE means cross-entropy loss. Following CIC [49], we quantize the ab color space into bins with grid size 10 and keep the $D = 313$ values which are in-gamut, *i.e.*, $\boldsymbol{Q}(I_{\text{coarse}}), \boldsymbol{Q}(I) \in \mathcal{R}^{64\times64\times313}$:

$$\boldsymbol{Q}(I)_{i,j} = \text{argmin}_{k\in\{0,1,\ldots,D\}} \left\| \left[\frac{I_{i,j}}{10}\right] - C_k \right\|^2, \tag{6}$$

where i, j indicates indices of x- and y-axis, C_k denotes ab values of color class k. The overall training objective of CIA is as follows and we simply set $\lambda = 1$ and $\lambda_{\text{ab}} = 0.1$.

$$\mathcal{L}_{coarse} = \|I_{\text{coarse}} - I\|^2 + \lambda \cdot \mathcal{L}_{\text{per}}(I_{\text{coarse}}, I) + \lambda_{\text{ab}} \cdot \mathcal{L}_{\text{ab}}(I_{\text{coarse}}, I). \tag{7}$$

3.3.2 Fine colorization. CLIP features cannot represent local informations of image in a finer granularity, thus the coarse colorization results only provide rough color-instance correspondence. Then in the fine stage, on the basis of the semi-colorized latent $\hat{z}$, we use the diffusion model to match the color words and instances precisely. The fine-stage objective function is

$$\mathcal{L}_{fine} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I),t}[\|\epsilon - \epsilon_\theta(z_t, c_{text}, \hat{z}, z_{edge}, f_s, t)\|^2], \tag{8}$$

where z_{edge} and f_s denotes edge and semantic map feature conditions extracted from I_{gray}, which will be introduced in Sec. 3.4.

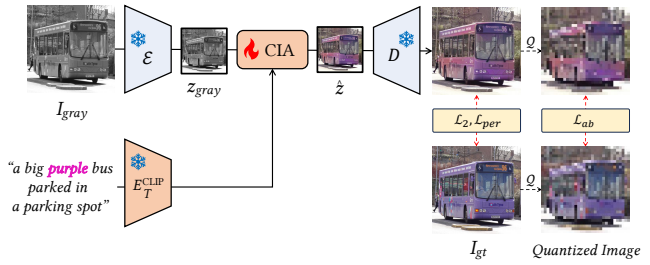

Figure 3: Illustration of Colorful Information Adaptor (CIA). We utilize six ResBlocks to merge grayscale content and color information hierarchically, following a parameterized version of AdaIN [16]. We utilized chrominance loss in quantized *LAB* color space for more colorful coarse results.

3.4 Consistency-aware Multi-level Condition

3.4.1 Low-level dual-branch condition insertion. As has been analyzed in Sec. 3.1, vanilla ControlNet has imbalanced feature granularity with the diffusion branch, which is especially harmful when dealing with gray image conditions, since grayscale images hold rich structure and semantic information than simple conditions like depth maps, and Canny edge maps [48]. Vanilla ControlNet do not have the ability to extract enough semantic features to precisely control the diffusion generation, leading to color overflow, color mismatch and structure distortion.

To balance the feature granularity between condition injection and diffusion generation, we propose a dual-branch feature extractor based on $\mathcal{E}$. Given a grayscale image I_{gray}, we utilize SAM [24] as a zero-shot edge detector to extract its instance-aware edge map I_{edge}, and encode it into the latent space $z_{edge} = \mathcal{E}(I_{edge})$. As mentioned before, we have inserted color information into z_{gray} with CIA, to obtain $\hat{z}$ with rich color priors. We concatenate $\hat{z}$ and z_{edge}, and fuse them using a single convolution layer. We use $\tilde{z} = Conv([\hat{z}, z_{edge}])$ as the input of the ControlNet branch.

3.4.2 High-level semantic feature modulation. To provide high-level feature regularization to build more accurate correspondence between the color and the instances, we leverage Mask2Former [8] to predict semantic segmentation maps based on I_{gray}. As the semantic map assigns labels to each pixel, we treat each label as a text description and use CLIP text encoder E_T^{CLIP} to extract the feature of each description appear in semantic maps. Finally, we aggregate these text features into a standard spatial semantic feature $f_s \in \mathcal{R}^{c \times h \times w}$, where c represents the dimension of CLIP text features and $h \times w$ is the resolution of I_{gray}.

We design a semantic feature fusion block (SFF) to modulate the original features f_{ori} on the skip connection of U-Net with f_s, as illustrated in Fig. 4. SFF adapts SPADE [30] with minor changes, and uses two submodules F_1 and F_2 to process and fuse the features in a "scale-shift" manner:

$$\hat{f}_{ori} = f_{ori} + LN(f_{ori}) \cdot F_1(f_s) + F_2(f_s), \tag{9}$$

where LN is the Layer Normalization. F_1 and F_2 denote two mapping network to wrap f_s into scale-shift parameters, with some convolution layers, activation functions and downsampling layers.

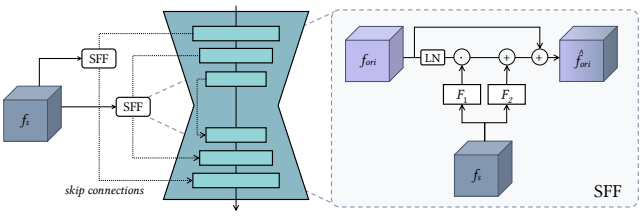

Figure 4: Illustration of high-level feature regularization

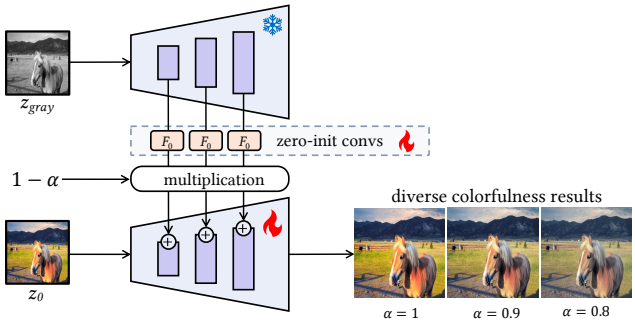

Figure 5: The structure of Colorfulness Controllable (COCO) module which merges the feature space of two VAE decoder space. We extract the middle features during the decoding of gray latents, wrap them with zero-initialized convolution layers and add them with the middle features of the decoding of colorized latents.

3.5 Colorfulness Controllable Colorization

To generate colorization results with varying color richness according with the user preference, we present colorfulness controllable colorization decoder $\mathcal{D}_{\text{COCO}}$ (COCO-decoder), based on the VAE decoder $\mathcal{D}$ with a scaling factor α. As shown in Fig. 5, $\mathcal{D}_{\text{COCO}}$ maintains a fixed decoder $\mathcal{D}$, a trainable decoder $\hat{\mathcal{D}}$ and trainable zero-initialized convolution layers F_0. We feed the gray image latent feature z_{gray} into $\mathcal{D}$ and get a set of middle features $\{d_i\}$ as the structure guidance, where i is the layer index. Correspondingly, we feed the diffusion output z_0 in the fine colorization stage into $\hat{\mathcal{D}}$, and obtain the corresponding features $\{\hat{d}_i\}$. In $\mathcal{D}_{\text{COCO}}$, we update $\hat{d}_i$ with d_i by

$$\hat{d}_i \leftarrow \hat{d}_i + (1-\alpha)F_0(d_i). \tag{10}$$

During the training of $\mathcal{D}_{\text{COCO}}$, we set $\alpha = 0$ and optimize

$$\mathcal{L}_{COCO} = \mathbb{E}_I[\left\|\mathcal{D}_{\text{COCO}}(z_0, z_{gray}, \alpha) - I)\right\|^2]. \tag{11}$$

Note that we use zero-initialized convolution layers to warm up the middle features during injection, preserving the capability of the pre-trained VAE decoder.

In the inference phase, we can use α to control the injection of gray information, *i.e.*, we can get diverse colorfulness results ranging from fantastic and gorgeous (high α), bright and realistic (middle α) to grayish and vintage (low α), enabling users to choose the best colorized result according to their preference.

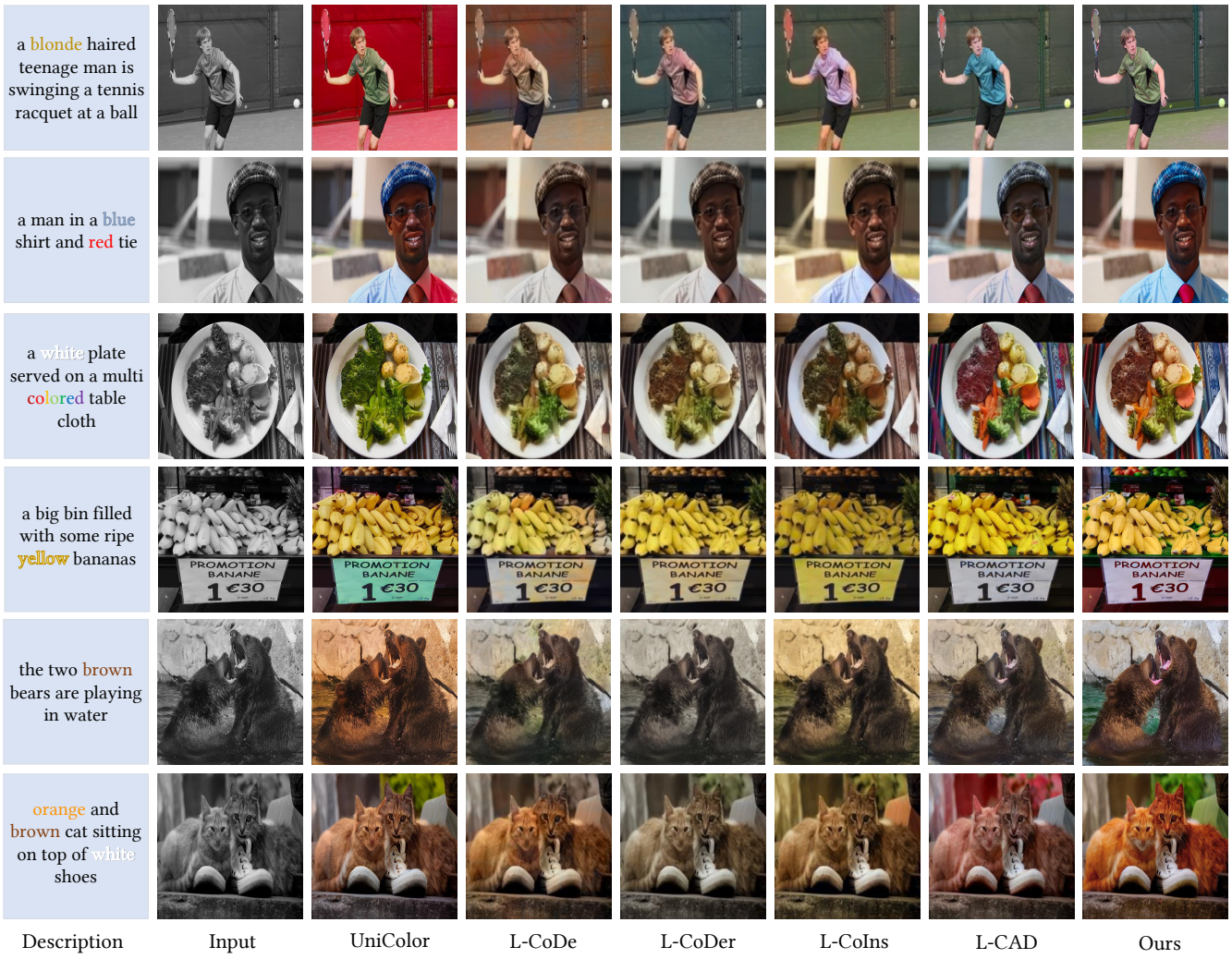

Figure 6: Qualitative comparison results of our methods and other language-based methods. Our colorized images enjoy a higher saturation and more plausible visual effect. Zoom in for better visualization.

4 Experiments

4.1 Implementation Details

Training. COCO-LC is trained on a single NVIDIA RTX 4090 GPU for 166k iterations with a batch size of 30. We adopt Deepspeed [33] to save GPU memory during training. We use AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is set to 10^{-5}. **CIA** is trained with 6 NVIDIA RTX 2080Ti GPUs for 80k iterations with a batch size of 60. The group number in GN is set to 64.

Inference. We use a single NVIDIA RTX 2080Ti GPU for inference. All testing images are resized to 512×512 with bilinear interpolation. Besides, we transform the output image to LAB space and replace its L channel with that of I_{gray}, to maintain the structure consistency. Our method will generate colorized results of the input grayscale images based on users' text description. If users don't provide any text descriptions, we utilize BLIP [26] to get a standard text description and do colorization automatically.

4.2 Evaluation

Training data. We conduct our experiments on language-based colorization datasets proposed by L-CoDe [42] and L-CoIns [7]: (i) the extended COCO-Stuff dataset, which is built upon the COCO-Stuff dataset [3] by discarding unqualified samples for the colorization task. We further filter out some black and white photos, remains 56k and 3,520 pairs of test images and prompts for training and testing respectively; (ii) multi-instances dataset which includes multiple instances with different visual characteristics within a single image, including 65K and 12714 pairs of test images and prompts for training and testing respectively.

Evaluate data. Apart from the validation set of the above two training datasets which totally contain 16,234 pairs of images and prompts, we use the first 5k images in ImageNet validation dataset, *i.e.* ImageNet-val5k. For the extended COCO-Stuff dataset and multi-instances dataset, each image is accompanied by several corresponding language descriptions. For ImageNet-val5k, we use BLIP [26]

to generate a default text description of the color image to evaluate our method.

Evaluation metrics. Following UniColor [17], we use Frechet Inception Distance (FID) [15] and colorfulness [13] to quantitatively evaluate the quality of our colorization results. We also utilize Δcolorfulness that computes the absolute difference of ground truth and colorization results to indicate realism in advance. Moreover, to evaluate prompt alignments, we calculate CLIP similarity score [31] between the prompts and colorization results. As image colorization is an ill-posed problem which may have multiple reasonable solutions, we do not adopt Peak Signal-to-Noise Ratio (PSNR) [18], Structural Similarity Index Measure (SSIM) [40] or Learned Perceptual Image Patch Similarity (LPIPS) [50] for precisely match pairs of colorized result and ground truth.

4.3 Comparisons

We make comparisons with both automatic colorization methods and language-based coolorization methods. For automatic colorization methods, we compare our method with CIC [49] (CNN-based), BigColor [23] (GAN-based), DDColor [22] and CT2 [41] (Transformer-based). For language-based colorization methods, we compare our methods with UniColor [17], L-CoDe [42], L-CoDer [5], L-CoIns [7] and L-CAD [6].

Qualitative comparisons As shown in Fig. 6, we make comparisons with previous language-based colorization methods. Our method reduces color overflow better and generates more bright and colorful results. While L-CoDe [42], L-CoDer [5], L-CoIns [7] may suffer from color overflow of yellow sign on the forth row, or struggle with color under-saturation, UniColor [17] produces unreasonable colors (*e.g.*the green meat in the third row) when conditioned by complex text descriptions and inaccurate colors (*e.g.*brown ties in the second row should be red in the text description). L-CAD [6] can generate globally natural images, but it fails to reduce color overflow in some local area, such as the red tennis racket in the first row, the red color of tie overflows to the white shirt in the second row, and the unreasonable magenta knife in the third row. Additionally, although L-CAD utilizes diffusion model to generate high-realistic results, it still suffers from grayish and unnatural results, which can be seen obviously in the last row. The green leaf and trees behind two cats are recognized by our model and colorized to proper colors, while other methods can only generate gray or inproper colors.

Quantitative comparisons As presented in Table 1, we present three different variants with different scaling factors to generate different styles of colors, ranging from fantastic ($\alpha = 1$), realism ($\alpha = 0.9$) to vintage ($\alpha = 0.8$). We make comparison with four unconditional method, CIC [49], BigColor [23], DDColor [22] and CT2 [41] , including CNN-based, Transformer-based and GAN-based methods. The comparison between other five language-based colorization methods: Unicolor [17], L-CoDe [42], L-CoDer [5], L-CoIns [7] and L-CAD [6], indicating our method can produce more plausible and accurate results. It is noticable that our method can not only generate high-realistic colorization results with the best FID and Δcolorfulness, but also generate the most colorful results with the highest colorfulness, satisfy the appetites of different users.

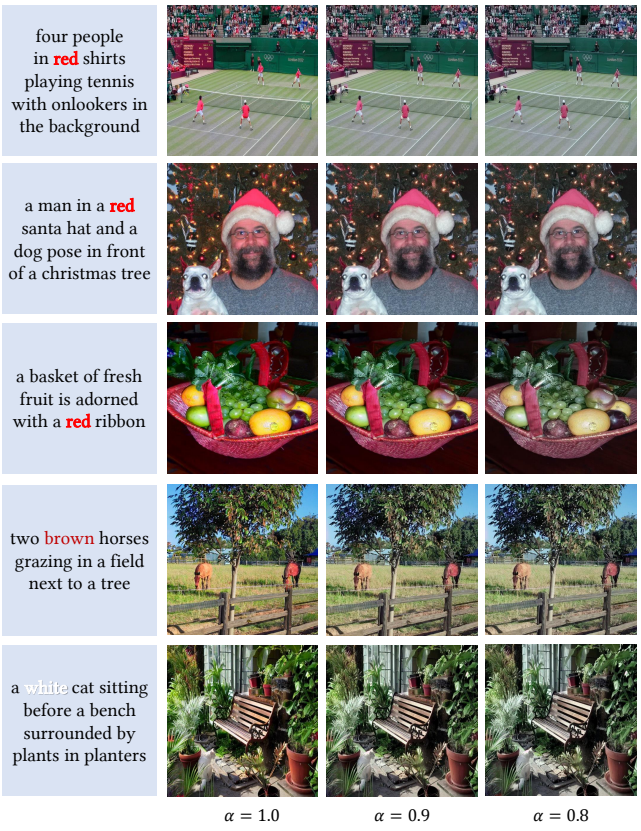

Figure 7: Qualitative results of our Colorfulness Controllable Decoder (COCO-Decoder) with scaling factor $\alpha \in \{0.8, 0.9, 1.0\}$. Our results enjoy different color styles ranging from fantastic ($\alpha = 1.0$), realistic ($\alpha = 0.9$) to vintage ($\alpha = 0.8$). We allow users to control colorfulness in a simple and flexible way.

4.4 User Study

We further conduct user studies to evaluate the subjective perception of human observers. We invite 46 volunteers to answer 10 questions, each questions contains a text description and eight colorized results of previous methods and our results. We encourage participants to evaluate those colorized results from the following three aspects: (1) consistency with text descriptions; (2) realism of images; (3) personal preference. We present our results of different colorfulness ranging from fantistic, realistic, to vintage at the same time to evaluate the robustness of our COCO-Decoder. The statistics results are presented as Fig. 8, which shows our method is prefered by most users.

4.5 Ablation Study and Discussion

We conduct other four baselines to demonstrate effectiveness of our coarse-to-fine framework and multi-level condition injection. The colorization results can be seen in Fig. 9.

vanilla ControlNet. We train a vanilla ControlNet without any design. We experimentally find that vanilla ControlNet will lead to color incomplete and color overflow. As shown in Fig. 9, tiny yellow spots appear on the bus body, and the number on the top of the

Table 1: Quantitative evaluation of different image colorization methods.

Dataset	Extended COCO-Stuff				Multi-instances				ImageNet5k-val			
Metrics	FID↓	Colorfulness↑	ΔColorfulness↓	CLIP Score↑	FID↓	Colorfulness↑	ΔColorfulness↓	CLIP Score↑	FID↓	Colorfulness↑	ΔColorfulness↓	CLIP Score↑
CIC [49]	29.05	29.86	13.59	-	17.55	26.66	14.65	-	13.35	25.95	10.69	-
BigColor [23]	27.21	48.32	4.87	-	11.85	47.36	6.05	-	12.18	44.57	7.93	-
DDColor [22]	14.94	40.85	2.6	-	4.89	43.5	2.19	-	5.56	42.35	5.71	-
CT2 [41]	18.21	25.31	18.14	-	10.64	41.02	0.29	-	7.51	39.56	2.92	-
UniColor [17]	16.03	41.4	2.05	26.43	19.32	33.53	7.78	26.42	10.63	34.98	1.66	28.4
L-CoDe [42]	31.21	29.48	13.97	29.78	25.08	25.18	15.15	27.59	19.56	24.23	12.41	27.66
L-CoDer [5]	31.78	29.52	13.93	29.98	22.75	26.45	13.88	27.53	16.73	22.90	13.74	27.93
L-CoIns [7]	33.67	25.31	18.14	30.06	23.13	30.65	9.68	27.61	21.03	33.53	3.11	28.57
L-CAD [6]	**12.75**	**43.8**	**0.35**	**30.7**	7.47	33.27	8.04	29.34	**9.3**	36.34	**0.3**	29.02
ours-fantastic	15.26	36.63	6.82	**30.7**	**6.37**	**35**	**6.31**	28.96	9.57	**37.4**	0.76	**29.32**
ours-realism	15.4	34.96	8.49	30.65	6.46	30.87	10.44	**29.42**	9.32	33.41	3.23	29.29
ours-vintage	15.34	32.13	11.32	30.37	6.57	29.83	11.48	28.98	9.45	30.29	6.35	29.3

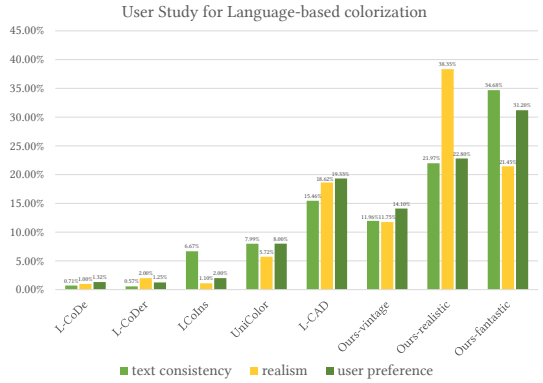

Figure 8: Quantitative results of user study.

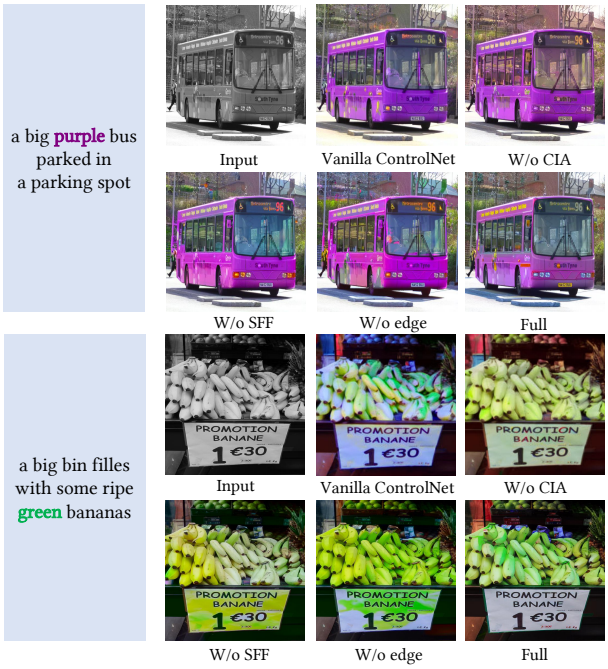

Figure 9: Qualitative results of ablation experiments.

bus is not completely colorized. The cloth of bus driver becomes purple due to the color overflow.

Colorful Information Adaptor. We remove CIA which insert CLIP color priors to grayscale latents. Without CIA, it's more difficult to build the accurate correspondence between color words and gray instances, leading to desaturation and misalignment of text descriptions. In Fig. 9, the purple of the bus overflows to the sky and the front glass, and the banana fails to become "green" precisely.

Semantic Feature Fusion. We disable the insertion of the spatial semantic feature. As a result, obvious color overflow occurs in the colorized images. In Fig. 9, the pole and person behind the bus becomes purple and the paper sign under the bananas becomes green.

Low-level Edge Condition. We disable the low-level edge condition, leading to color incompletion and color overflow, such as gray artifacts on purple bus and green smears on white sign.

COCO-Decoder. We present some results of different color styles by our colorfulness controllable decoder with different scaling factor, as shown at Fig. 7. Our method provides a user-friendly way of controllable color richness to generate diverse colors ranging from bland to gorgeous. Please refer to our supplementary materials for more visualization results.

In summary, we propose a coarse-to-fine framework that use CIA to insert rich color priors conditioned by language prompts in the coarse stage. On the basis of this semi-colorized result, we use high-level spatial semantic features and low-level edge latent codes to constrain colors to the correct area spatially. Disabling any one of them will lead to obvious color overflow.

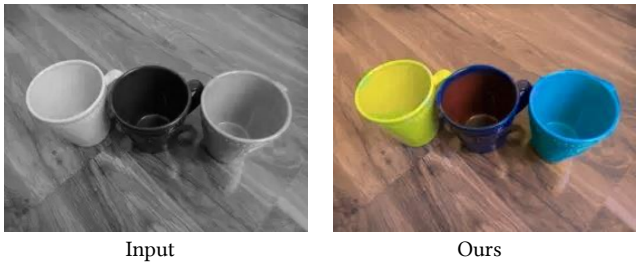

Figure 10: Failure case of our method when dealing with multiple color words in a text prompt. Prompt: "the blue cup, the yellow cup and the cyan cup". The middle cup is incorrectly rendered red.

5 Conclusion and Discussion

In this paper, we present **COCO-LC**, a novel coarse-to-fine framework that achieves COlorfulness COntrollable Language-based Colorization. We design a multi-level condition to reduce color overflow and COCO-Decoder to generate colorized results with diverse color styles flexibly. Extensive experiments demonstrate the superiority of COCO-LC over state-of-the-art image colorization methods in accurate, realistic and controllable colorization.

Limitations and Future Work. Stable Diffusion utilizes CLIP to align text and image domain, which struggles with complex text descriptions. When there are multiple colors and instances, it is hard to find accurate correspondence, leading to color-instance mismatch, as shown in Fig. 10. The color of the middle blue cup turns red unexpectly. In our future work, we would like to adopt more powerful cross-modality models and generative backbones to enhance the robustness of colorization.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62332010, and in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

References

[1] Jason Antic. 2019. *DeOldify: A Deep Learning Based Project for Colorizing and Restoring Old Images (and Video!).* https://github.com/jantic/DeOldify

[2] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Cold diffusion: Inverting arbitrary image transforms without noise. In *Advances in Neural Information Processing Systems.*

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. Coco-stuff: Thing and stuff classes in context. In *IEEE/CVF Int'l Conferece on Computer Vision and Pattern Recognition.*

[4] Yun Cao, Zhiming Zhou, Weinan Zhang, and Yong Yu. 2017. Unsupervised diverse colorization via generative adversarial networks. In *Machine Learning and Knowledge Discovery in Databases: European Conf., ECML PKDD 2017.*

[5] Zheng Chang, Shuchen Weng, Yu Li, Si Li, and Boxin Shi. 2022. L-CoDer: Language-based Colorization with Color-object Decoupling Transformer. In *European Conf. on Computer Vision.*

[6] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. 2023. L-CAD: Language-based Colorization with Any-level Descriptions using Diffusion Priors. In *Advances in Neural Information Processing Systems.*

[7] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. 2023. L-CoIns: Language-based Colorization with Instance Awareness. In *IEEE/CVF Int'l Conferece on Computer Vision and Pattern Recognition.*

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Int'l Conferece on Computer Vision and Pattern Recognition.*

[9] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. 2015. Deep colorization. In *IEEE/CVF Int'l Conf. on Computer Vision.*

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805* (2018).

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems.*

[12] Sergio Guadarrama, Ryan Dahl, David Bieber, Mohammad Norouzi, Jonathon Shlens, and Kevin Murphy. 2017. Pixcolor: Pixel recursive colorization. *arXiv:1705.07208* (2017).

[13] David Hasler and Sabine E Suesstrunk. 2003. Measuring colorfulness in natural images. In *Human Vision and Electronic Imaging VIII.*

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE/CVF Int'l Conferece on Computer Vision and Pattern Recognition.*

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems.*

[16] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE/CVF Int'l Conf. on Computer Vision.*

[17] Zhitong Huang, Nanxuan Zhao, and Jing Liao. 2022. Unicolor: A unified framework for multi-modal colorization with transformer. *ACM Transactions on Graphics* (2022).

[18] Quan Huynh-Thu and Mohammed Ghanbari. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters* (2008).

[19] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2016. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics* 35 (2016).

[20] Xiaozhong Ji, Boyuan Jiang, Donghao Luo, Guangpin Tao, Wenqing Chu, Zhifeng Xie, Chengjie Wang, and Ying Tai. 2022. ColorFormer: Image colorization via color memory assisted hybrid-attention transformer. In *European Conf. on Computer Vision.*

[21] Xin Jin, Zhonglan Li, Ke Liu, Dongqing Zou, Xiaodong Li, Xingfan Zhu, Ziyin Zhou, Qilong Sun, and Qingyu Liu. 2021. Focusing on Persons: Colorizing Old Images Learning from Modern Historical Movies. In *ACM Int'l Conf. on Multimedia.*

[22] Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, and Xuansong Xie. 2023. DDColor: Towards Photo-Realistic Image Colorization via Dual Decoders. In *IEEE/CVF Int'l Conf. on Computer Vision.*

[23] Geonung Kim, Kyoungkook Kang, Seongtae Kim, Hwayoon Lee, Sehoon Kim, Jonghyun Kim, Seung-Hwan Baek, and Sunghyun Cho. 2022. BigColor: Colorization using a generative color prior for natural images. In *European Conf. on Computer Vision.*

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *IEEE/CVF Int'l Conf. on Computer Vision.*

[25] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. 2021. Colorization Transformer. In *Proc. Int'l Conf. Learning Representations.*

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *IEEE Int'l Conf. on Machine Learning.*

[27] Zekun Li, Zhengyang Geng, Zhao Kang, Wenyu Chen, and Yibo Yang. 2022. Eliminating Gradient Conflict in Reference-based Line-Art Colorization. In *European Conf. on Computer Vision.*

[28] Zhexin Liang, Zhaochen Li, Shangchen Zhou, Chongyi Li, and Chen Change Loy. 2024. Control Color: Multimodal Diffusion-based Interactive Image Colorization. *arXiv:2402.10855* (2024).

[29] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. 2023. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv:2308.15070* (2023).

[30] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic Image Synthesis With Spatially-Adaptive Normalization. In *IEEE/CVF Int'l Conferece on Computer Vision and Pattern Recognition.*

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *IEEE Int'l Conf. on Machine Learning.*

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125* (2022).

[33] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 3505–3506.

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Int'l Conferece on Computer Vision and Pattern Recognition.*

[35] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Int'l Conferece on Computer Vision and Pattern Recognition.*

[36] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. 2020. Instance-aware Image Colorization. In *IEEE/CVF Int'l Conferece on Computer Vision and Pattern Recognition.*

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems.*

[38] Patricia Vitoria, Lara Raad, and Coloma Ballester. 2020. ChromaGAN: Adversarial Picture Colorization with Semantic Class Distribution. In *IEEE/CVF Winter Conf. on Applications of Computer Vision.*

[39] Yi Wang, Menghan Xia, Lu Qi, Jing Shao, and Yu Qiao. 2022. PalGAN: Image Colorization with Palette Generative Adversarial Networks. In *European Conf. on Computer Vision.*

[40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* (2004).

[41] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. 2022. CT2: Colorization Transformer via Color Tokens. In *European Conf. on Computer Vision*.
[42] Shuchen Weng, Hao Wu, Zheng Chang, Jiajun Tang, Si Li, and Boxin Shi. 2022. L-CoDe: Language-based colorization using color-object decoupled conditions. In *AAAI Conf. on Artificial Intelligence*.
[43] Shukai Wu, Yuhang Yang, Shuchang Xu, Weiming Liu, Xiao Yan, and Sanyuan Zhang. 2023. FlexIcon: Flexible Icon Colorization via Guided Images and Palettes. In *ACM Int'l Conf. on Multimedia*.
[44] Yuxin Wu and Kaiming He. 2018. Group Normalization. In *European Conf. on Computer Vision*.
[45] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. 2023. Diffir: Efficient diffusion model for image restoration. In *IEEE/CVF Int'l Conf. on Computer Vision*.
[46] Jooyeol Yun, Sanghyeon Lee, Minho Park, and Jaegul Choo. 2023. iColoriT: Towards Propagating Local Hints to the Right Region in Interactive Colorization by Leveraging Vision Transformer. In *IEEE/CVF Winter Conf. on Applications of Computer Vision*.
[47] Nir Zabari, Aharon Azulay, Alexey Gorkor, Tavi Halperin, and Ohad Fried. 2023. Diffusing Colors: Image Colorization with Text Guided Diffusion. In *ACM SIGGPRAH Asia*.
[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF Int'l Conf. on Computer Vision*.
[49] Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *European Conf. on Computer Vision*.
[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Int'l Conferece on Computer Vision and Pattern Recognition*.
[51] Jiaojiao Zhao, Jungong Han, Ling Shao, and Cees GM Snoek. 2020. Pixelated semantic colorization. *Int'l Journal of Computer Vision* 128 (2020).