L-CoIns: Language-based Colorization with Instance Awareness

Zheng Chang[#1] Shuchen Weng[#2,3] Peixuan Zhang[1] Yu Li[4] Si Li[*1] Boxin Shi[2,3]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications

[2]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University

[3]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

[4]International Digital Economy Academy

{zhengchang98,pxzhang,lisi}@bupt.edu.cn, {shuchenweng,shiboxin}@pku.edu.cn, liyu@idea.edu.cn

Figure 1. Language-based colorization results given four different language descriptions, compared with ML2018 [29], L-CoDe [42], and L-CoDer [6]. **Top left:** For the description that has clear correspondences between color words and object words, our method correctly colorizes all corresponding regions. **Top right:** For the description that assigns distinct colors for every instance corresponding to the same object words, our model predicts the exact correspondence between the instance region and the color word. **Botton left:** For the description that includes unobserved correspondences between color words and object words, our method could adaptively parse the sentence and determine the correct semantics for colorization. **Bottom right:** For the description that is against the statistical correlation between luminance and color words, our method shows the robustness and colorize description-consistent results.

Abstract

Language-based colorization produces plausible colors consistent with the language description provided by the user. Recent studies introduce additional annotation to prevent color-object coupling and mismatch issues, but they still have difficulty in distinguishing instances corresponding to the same object words. In this paper, we propose a transformer-based framework to automatically aggregate similar image patches and achieve instance awareness without any additional knowledge. By applying our presented luminance augmentation and counter-color loss to break down the statistical correlation between luminance and color words, our model is driven to synthesize colors with better descriptive consistency. We further collect a dataset to provide distinctive visual characteristics and detailed language descriptions for multiple instances in the same image. Extensive experiments demonstrate our advantages of synthesizing visually pleasing and description-consistent results of instance-aware colorization.

1. Introduction

Image colorization aims to predict missing chromatic channels from a given grayscale image, which has been widely used in black-and-white image restoration, artistic creation, and image compression. Since there are multiple reasonable choices for the colorization result, an increasing amount of effort has focused on introducing user-friendly interactions to determine a unique solution, *e.g.*, user scribble [33, 51], and reference example [2, 16, 47]. In contrast to these visually-concrete conditions, the language descriptions have higher information density to flexibly represent high-level semantics, which empowers the colorization model to concrete visually-abstract user intention.

Language-based colorization aims to produce visually

[#] Equal contributions. * Corresponding author.

pleasing and description-consistent results guided by the user-provided caption. In such a task, the most crucial stage is to establish the correspondence between the colors in the language description and the regions in the image. Cross-modality feature fusion modules are designed in earlier methods [8, 29, 45, 56], but they are ineffective in generating satisfactory results on samples with fewer observed color-object correspondences and insufficient color descriptions. By introducing additionally annotated correspondences between object words and color words, remarkable improvements are observed on recently reported results on a wide variety of images [6, 42], but these methods still face challenges in distinguishing instances corresponding to the same object words (*e.g.*, the "woman" in Fig. 1 top/bottom right). While introducing additional external priors (*e.g.*, detection boxes [35]) is an alternative approach to achieve instance-aware colorization, it may not perform well on "out-of-distribution" scenarios [41].

In this paper, we propose **L**anguage-based **Co**lorization with **Ins**tance awareness (**L-CoIns**) to adaptively establish the correspondence between instance regions and color descriptions without additionally using external priors. L-CoIns considers an image as a composition of a number of groups with similar colors, hence adopting a grouping mechanism to automatically aggregate similar image patches for correctly identifying corresponding regions to be colorized (Fig. 1 top left, regions of women are correctly identified) and distinguishing instances corresponding to the same object words (Fig. 1 top right, corresponding colors are assigned to different instances) in an unsupervised manner. Our model is able to more flexibly assign colors for instances, even when correspondences never occur during training, as opposed to learning manually annotated multiple color-object correspondences (Fig. 1 bottom left, the correspondence between violet and shirt is unobserved). We propose the luminance augmentation and counter-color loss to break down the statistical correlation between luminance and color words so that L-CoIns could produce colorization results that are more consistent with the given language description (Fig. 1 bottom right, yellow and orange successfully colorize darker and brighter regions).

Our contribution could be summarized as follows:

- Without additionally annotating correspondences or external priors, we provide the grouping transformer to aggregate similar image patches and learn inter-group relations for instance-aware language colorization.
- We present the luminance augmentation and counter-color loss that stick the model to colorize according to the language description rather than the statistical correlation between luminance and color words.
- We collect a multi-instance dataset that offers miscellaneous cases with distinctive visual characteristics and detailed language descriptions for various instances within an image.

2. Related Works

Automatic colorization. Automatic colorization pursues to generate diverse, colorful, and plausible results in a data-driven manner without additional user-provided guidance. Chen *et al.* [9] build the first deep-learning based colorization model by using handcraft features. After that, researchers focus on designing adaptive feature extraction modules by introducing CNN [19, 25, 49]. Recently, the colorization model architecture gradually moves to transformer [15] by proposing novel attention modules [20, 24, 41]. In addition to exploring advanced generative models (*e.g.*, VAE [11], GAN [4, 38], INN [1]) for creating vivid colorization results, researchers also pay attention to adopting external priors (*e.g.*, pretrained generative model [22, 43], categories [38], and semantic segmentation [52, 53]) to obtain high-level semantic understanding of images and further improve the colorization fidelity. Especially, InstColor [35] uses a pretrained detection model to predict bounding boxes of instances so that multiple objects could be separately colorized. However, taking external priors relies heavily on the performance of upstream models [41]. To take a step towards general colorization scenarios, recent works introduce predefined priors (*e.g.*, palette histogram [39] and superpixel [44]) to guide colorization by specifically designed supervisory signals. In this paper, we further explore how to perform instance-aware colorization without requiring additional external knowledge.

Language-based colorization. Given a user-provided language description that contains objects and their respective colors, language-based colorization aims to produce appropriate colorization results consistent with the description. Manjunatha *et al.* [29] design novel feature-wise affine transformations to inject language condition into image features and propose the first language-based colorization method. Similar approach is adopted by Chen *et al.* [8], which fuses image and language features spatially by a recurrent attentive model. Inspired by automatic colorization methods [53], Xie *et al.* [45] introduce external priors to capture high-level image semantic by learning semantic segmentation as a side task. Recently, UniColor [17] designs the first unified framework to support colorization with multi-modal interactions (*e.g.*, language, scribble, and exemplar). Although these methods have made great progress in generating vivid colorization results, it is difficult to correctly inject color words into corresponding image regions due to the semantic chasm between image and language. As alternative solutions, L-CoDe [42] and L-CoDer [6] employ additionally annotated correspondence between object words and color words to decouple language description into object space and color space so that they

could assign image regions with specific color words. However, these approaches make the vocabulary and the flexibility of description limited by the annotated correspondence, which motivates us to come up with a new solution that adaptively learns the correspondence between instances regions and color words.

Vision transformer. Transformer [37] is firstly demonstrated effective in natural language processing with the multi-head attention mechanism to model global token-to-token relationships. Recently, researchers have successfully developed vision transformer for a wide range of vision applications, *e.g.*, image classification [15], object detection [5, 55], and semantic segmentation [34, 54]. Great efforts have also been made to adapt vision transformer models to low-level vision problems, *e.g.*, inpainting [26, 28], super resolution [7, 27], and colorization [20, 24, 41]. Observing that the unique feature fusion mechanism of transformer is naturally applicable for cross-modality tasks, *e.g.*, text-to-image generation [14, 32], referring segmentation [13, 46, 48], and visual grounding [10, 31], we also build our language-based colorization solution based on the transformer architecture as the previous work [6].

3. Methodology

This section provides a brief overview of L-CoIns before going into further details that elaborate designs of the modules.

3.1. Framework

We show the pipeline of L-CoIns in Fig. 2, which could be divided into the following four steps: *(i)* **Hiding illuminance cues.** To prevent the colorization model from assigning the color to instances based on the statistical correlation between luminance and color words rather than understanding the language description, we randomly transform the illuminance with luminance augmentation before feeding grayscale images into the colorization model. *(ii)* **Unifying cross-modality conditions.** In addition to introducing multiple learnable vectors as group tokens, we also employ separate transformers to encode grayscale images and language descriptions into tokens so that they could be mapped into the same representation to bridge the semantic chasm. *(iii)* **Learning inter-token relationship.** To enable global interaction between tokens and improve understanding of the relationship between tokens, a grouping transformer is proposed to inject color features into image tokens, aggregate image features into group tokens, and finally correctly colorize the corresponding instance. *(iv)* **Optimizing colorization error.** Beyond optimizing the chromatic error in a regression manner, we further design a counter-color loss as the grouping supervisory signal to reduce grouping error, allowing group tokens to separate instances automatically.

3.2. Luminance Augmentation

In contrast to automatic colorization methods that are allowed to infer the most common colors from the input luminance, languaged-based colorization methods are expected to synthesize specific colors under the guidance of the user-provided language description. However, the dataset randomly collected from the Internet tends to have a long-tailed distribution for correlation between luminance and color words, *e.g.*, the yellow is more likely to be bright, whereas the orange is typically darker. As a result, some language-based colorization models use luminance to infer the corresponding color in the language description for each instance and further produce incorrect colorization results when descriptions are against the statistical correlation between luminance and colors, as shown in the bottom right of Fig. 1. Therefore, we design the luminance augmentation to randomly transform image luminance during the training stage and drive the model to colorize by understanding the description provided by the user.

First, we convert the original RGB image into HSV color space, which allows the hue to be independent of saturation and brightness. Considering that HSV color space is a conical geometry with red at $0°$, green at $120°$, blue at $240°$, and again red at $360°$ as its starting point, we could randomly rotate the hue to modify the relative luminance across instances as:

$$I_{\mathrm{r}} = [\mathrm{F}_{\mathrm{rotate}}(I_{\mathrm{o}}, \lambda)^{\mathrm{h}}, I_{\mathrm{o}}^{\mathrm{s}}, I_{\mathrm{o}}^{\mathrm{v}}], \tag{1}$$

where $\mathrm{F}_{\mathrm{rotate}}$ is the rotation operator, I_{o} is the original HSV color image, the superscripts h, s, and v mean the corresponding channel of hue, saturation, and brightness. $\lambda \in [-180, 180]$ is the angle of rotation.

We then convert I_{r} into LAB color space, and separate the luminance channel as a grayscale image I_{g}. To avoid the statistical correlation between absolute luminance and color words, we further adjust the global luminance with a random gamma correction as:

$$\hat{I}_{\mathrm{g}} = A I_{\mathrm{g}}^{\mathrm{F}_{\mathrm{inv}}(\gamma)}, \tag{2}$$

where we set constant A to 1 by default, randomly uniform sample γ in intervals $[1, 5]$, and build $\mathrm{F}_{\mathrm{inv}}(x)$ as a probabilistic reciprocal function that changes x to $1/x$ with the chance of 50%.

This strategy could preserve texture details while breaking down the statistical correlation between luminance and colors, which drives the model to understand language descriptions to optimize the colorization error. We show the augmented grayscale images in Fig. 3.

3.3. Tokens Embedding

After performing the luminance augmentation, grayscale images and language descriptions are separately encoded

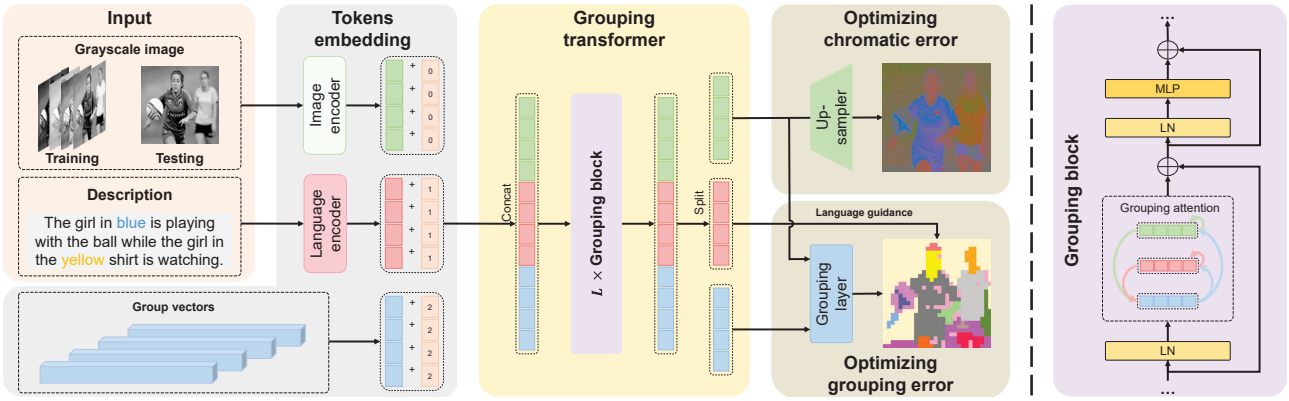

Figure 2. **Left:** The pipeline of L-CoIns. We adopt the luminance augmentation to break down the correlation between luminance and color words in the training stage firstly (Sec. 3.2). Next, we obtain multi-modal tokens by introducing learnable group vectors and encoding input image and language description separately (Sec. 3.3). After that, all tokens are concatenated and fed into global-interactive grouping transformers to extract high-level semantic features (Sec. 3.4). Finally, the upsampler (Sec. 3.5) and the grouping layer (Sec. 3.6) are employed to optimize the chromatic error and the grouping error (Sec. 3.7), respectively. **Right:** The structure of grouping block in the grouping transformer.

Figure 3. Examples of the luminance augmentation. In addition to enlarging (first column) or reversing (second column) the relative luminance between instances, the luminance augmentation could also increase (third column) or decrease (fourth column) the global luminance.

and multiple learnable vectors are introduced as multi-modal tokens. All tokens share the same representation to make global interaction in the next step and we will introduce these tokens one by one.

Image tokens. We repeat the augmented grayscale image $\hat{I}_{\mathrm{g}} \in \mathbb{R}^{H\times W}$ into a pseudo-color image $I_{\mathrm{c}} \in \mathbb{R}^{H\times W\times 3}$, where H and W are the image size. It is then split into patch sequence as $I_{\mathrm{pat}} = [I^1_{\mathrm{pat}}, \ldots, I^{N_{\mathrm{I}}}_{\mathrm{pat}}] \in \mathbb{R}^{N_{\mathrm{I}}\times P^2\times 3}$, where P is the patch size and $N_{\mathrm{I}} = HW/P^2$. We adopt a standard ViT [15] to encode patch sequence, which captures the long-range dependency in the sequence and generates image tokens $T_{\mathrm{img}} = [T^1_{\mathrm{img}}, \ldots, T^{N_{\mathrm{I}}}_{\mathrm{img}}] \in \mathbb{R}^{N_{\mathrm{I}}\times C_{\mathrm{T}}}$. Here we denote C_{T} as the channel number.

Language tokens. The language description is naturally a sequence, therefore the typical Transformer (*i.e.*, BERT [12]) could be applied to generate language tokens $T_{\mathrm{lag}} = [T^1_{\mathrm{lag}}, \ldots, T^{N_{\mathrm{L}}}_{\mathrm{lag}}] \in \mathbb{R}^{N_{\mathrm{L}}\times C_{\mathrm{T}}}$, where N_{L} is the sequence length. Thanks to the pretrained dictionary of BERT that includes more than 20K tokens, we could handle the isolated words that never appear in the training set.

Group tokens. We introduce N_{G} learnable vectors as group tokens to aggregate similar image patches and adaptively present distinct instances, which are denoted as $T_{\mathrm{grp}} = [T^1_{\mathrm{grp}}, \ldots, T^{N_{\mathrm{G}}}_{\mathrm{grp}}] \in \mathbb{R}^{N_{\mathrm{G}}\times C_{\mathrm{T}}}$.

After obtaining multi-modal tokens, we separately use modal-type embedding vectors [23] to distinguish modalities of image, language, and group, denoted as $T'_{\mathrm{img}}, T'_{\mathrm{lag}}, T'_{\mathrm{grp}}$. Then, these modal-type embedding vectors are added to corresponding tokens as:

$$\hat{T}_i = T_i + T'_i. \qquad i \in \{\mathrm{img}, \mathrm{lag}, \mathrm{grp}\} \tag{3}$$

3.4. Grouping Transformer

To learn inter-token relationships and extract high-level semantic information, we propose the grouping transformer equipped with the grouping attention, which uses group tokens as a medium to achieve bidirectional interaction between image and language tokens. In this way, different instances are gradually distinguished by group tokens, image tokens are progressively colorized, and language tokens evolve in a coarse-to-fine representation manner as the network grows deeper.

We propose the grouping transformer to learn the inter-group relationship between tokens, which is a stack of L grouping blocks. The grouping block is built based on the standard transformer block, where the multi-head attention is replaced with the novel grouping attention (GA). Given $[Z^l] = [Z^l_{\mathrm{img}}; Z^l_{\mathrm{lag}}; Z^l_{\mathrm{grp}}] \in \mathbb{R}^{(N_{\mathrm{I}}+N_{\mathrm{L}}+N_{\mathrm{G}})\times C^l_{\mathrm{Z}}}$ as the input of l-th grouping block and C^l_{Z} as the corresponding channel number, we formulate the process of the grouping block as:

$$[\hat{Z}^l] = \mathrm{GA}\ (\mathrm{LN}([Z^l])) + [Z^l], \quad l \in \{1, \ldots, L\} \tag{4}$$

$$[Z^{l+1}] = \mathrm{MLP}(\mathrm{LN}([\hat{Z}^l])) + [\hat{Z}^l], \quad l \in \{1, \ldots, L\} \tag{5}$$

where LN is the LayerNorm layer and $[\hat{T}_{\mathrm{img}}; \hat{T}_{\mathrm{lag}}; \hat{T}_{\mathrm{grp}}]$ is used as the initial $[Z^1]$.

After projecting $[Z^l]$ into query, key, and value feature space separately by fully connected layers, we organize the grouping attention as the two-type loops: *(i)* Inner loop: This loop aims at obtaining a deeper understanding of the given image and language feature as well as the acquisition of deeper semantic features. *(ii)* Outer loop: This loop considers group tokens as the medium to flow the semantic information between image and language features bidirectionally. As group tokens are the only means of interacting between images and language features in the grouping transformer, they are pushed to understand both semantics for optimizing the colorization error. Denoting $i \in \{\text{img}, \text{lag}\}$ as the modality of tokens, $h \in \{1, \dots, N_\text{H}\}$ as the index of attention heads, and $\frac{1}{\sqrt{d_\text{k}}}$ as the scaling factor, the inner loop could be achieved as self-attention:

$$Z^{\text{in}}_{i,h} = \text{F}_{\text{softmax}}(\frac{Z^{\text{qry}}_{i,h}(Z^{\text{key}}_{i,h})^\top}{\sqrt{d_\text{k}}})Z^{\text{val}}_{i,h}, \tag{6}$$

and the outer loop is represented as cross-attention:

$$Z^{\text{out}}_{\text{grp},h} = \text{F}_{\text{softmax}}(\frac{Z^{\text{qry}}_{\text{grp},h}(Z^{\text{key}}_{i,h})^\top}{\sqrt{d_\text{k}}})Z^{\text{val}}_{i,h}, \tag{7}$$

$$Z^{\text{out}}_{i,h} = \text{F}_{\text{softmax}}(\frac{Z^{\text{qry}}_{i,h}(Z^{\text{key}}_{\text{grp},h})^\top}{\sqrt{d_\text{k}}})Z^{\text{val}}_{\text{grp},h}. \tag{8}$$

We finally concatenate tokens of the same modality and project them into deeper feature space:

$$[\hat{Z}_*] = \text{F}_{\text{concat}}([Z^{\text{in}}_{*,1}\ Z^{\text{out}}_{*,1}\ \dots\ Z^{\text{in}}_{*,N_\text{H}}\ Z^{\text{out}}_{*,N_\text{H}}])W^{\text{proj}}, \tag{9}$$

where $*$ represents the label from $\{\text{img}, \text{lag}, \text{grp}\}$ and W^{proj} is a learnable parameter matrix. We omit superscript l from Eq. (6) to Eq. (9) for simplicity.

3.5. Upsampler

The upsampler is used to convert the colorized image tokens Z^{L+1}_{img} generated by the grouping transformer into chrominance channels in preparation for optimizing the chromatic error. To be more specific, we first reshape the image token sequence into the spatial resolution of $\sqrt{N_\text{I}} \times \sqrt{N_\text{I}}$, and then feed them into a stack of transposed convolutions to upsample and output the predicted chromatic channels:

$$\hat{I}_{\text{ab}} = \text{F}_{\text{up}}(\text{F}_{\text{rsp}}(Z^{L+1}_{\text{img}})), \tag{10}$$

where F_{up} means convolutional layers and F_{rsp} is the reshape operator.

3.6. Grouping Layer

As a preparation for optimizing grouping error, we explicitly map image tokens into their corresponding group tokens in a hard assignment manner [46]. Specifically, we calculate the similarity matrix $\bar{A} \in \mathbb{R}^{N_\text{G} \times N_\text{I}}$ between all image tokens and group tokens at the finest grain (denoted as Z^{L+1}_{img} and Z^{L+1}_{grp}), and then only retain the most similar group token for each image token:

$$\hat{A} = \text{F}_{\text{argmax}}(\bar{A}) + \bar{A} - \text{F}_{\text{sg}}(\bar{A}), \tag{11}$$

where F_{argmax} is the function to set non-maximum elements to 0 for each column and F_{sg} is the stop gradient operator to perform straight-through trick [36] and make the gradient differentiable. We further merge image tokens corresponding to the same group in a weighted sum manner to concrete the instance representation:

$$\bar{Z}^{L+1}_{\text{grp},i} = Z^{L+1}_{\text{grp},i} + \frac{\sum_{j=1}^{N_\text{I}} \hat{A}_{i,j} W_\text{v} Z^{L+1}_{\text{img},j}}{\sum_{j=1}^{N_\text{I}} \hat{A}_{i,j}}, \tag{12}$$

where W_v is a learnable matrix, the subscript i, j means i-th group token and j-th image token.

3.7. Loss Function

Following previous methods [6, 42], a regression approach could be used to optimize the chromatic error. Therefore, we use a smooth-ℓ_1 loss with $\delta = 1$ to supervise predicted chromatic channels as:

$$\begin{aligned} L_\delta = &\frac{1}{N_\text{p}} \sum \frac{1}{2}(\hat{I}_{\text{ab}} - I_{\text{ab}})^2 \mathbb{1}_{\{|\hat{I}_{\text{ab}} - I_{\text{ab}}| < \delta\}} + \\ &\frac{1}{N_\text{p}} \sum \delta(|\hat{I}_{\text{ab}} - I_{\text{ab}}| - \frac{1}{2}\delta) \mathbb{1}_{\{|\hat{I}_{\text{ab}} - I_{\text{ab}}| \geq \delta\}}, \end{aligned} \tag{13}$$

where N_p is the pixel number, I_{ab} is the ground truth of chromatic channels, and $\mathbb{1}$ means the tensor of one.

As there is no ground truth for instance separation, we further propose the counter-color loss to optimize the grouping error. We use two linear layers to map the latest language tokens Z^{L+1}_{lag} and group tokens $\bar{Z}^{L+1}_{\text{grp}}$ into the shared feature space as R_{lag} and R_{grp}, respectively. We then randomly replace the color words in the language description with another one to generate pseudo language tokens and group tokens, denoted as R'_{lag} and R'_{grp}. To measure the relevance between multi-modal tokens, we define the similarity function between language tokens and group tokens as:

$$\text{F}_{\text{sim}}(x, y) = \sigma\Big(\frac{1}{N_\text{L}} \frac{1}{N_\text{G}} \sum_{i=1}^{N_\text{L}} \sum_{j=1}^{N_\text{G}} \big(\text{F}_{\cos}(x_i, y_j)\big)\Big), \tag{14}$$

where σ is the sigmoid function, $x \in [R_{\text{lag}}, R'_{\text{lag}}]$, $y \in [R_{\text{grp}}, R'_{\text{grp}}]$, x_i and y_i represent the i-th language token and j-th group token, and $\text{F}_{\cos}$ is the function to calculate the cosine similarity. After that, the final counter-color loss could be written as:

$$\begin{aligned} L_{\text{ctr}} = &\log\Big(\text{F}_{\text{sim}}(R_{\text{lag}}, R_{\text{grp}})\big(1 - \text{F}_{\text{sim}}(R'_{\text{lag}}, R_{\text{grp}})\big)\Big) + \\ &\log\Big(\text{F}_{\text{sim}}(R'_{\text{lag}}, R'_{\text{grp}})\big(1 - \text{F}_{\text{sim}}(R_{\text{lag}}, R'_{\text{grp}})\big)\Big). \end{aligned} \tag{15}$$

Table 1. Quantitative experiment results of comparison and ablation. ↑ (↓) means higher (lower) is better. Best performances are highlighted in **bold**.

Comparison with state-of-the-art methods						
Method	Extended COCO-Stuff			Multi-instance		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
LBIE [8]	22.092	0.85197	0.265	21.928	0.85910	0.260
ML2018 [29]	21.055	0.85333	0.282	20.538	0.84954	0.294
Xie2018 [45]	21.407	0.84016	0.298	19.920	0.81712	0.321
L-CoDe [42]	24.965	0.91657	0.169	23.955	0.91283	0.180
L-CoDer [6]	25.504	0.91963	0.159	24.257	0.91302	0.175
Ours	**25.511**	**0.92104**	**0.157**	**24.823**	**0.91717**	**0.162**
Ablation study						
W/o GE	25.249	0.91832	0.163	24.230	0.90734	0.173
W/o GA	25.373	0.91913	0.160	24.576	0.91261	0.164
W/o LA	25.475	0.91965	0.161	24.382	0.91435	0.167
W/o CL	25.361	0.91866	0.158	24.475	0.91401	0.163

As a result of the counter-color loss, the group tokens are required to retrieve the corresponding language tokens, which forces the group tokens to focus more on the instance regions corresponding to the modified color words, thus requiring group tokens to separate the instances appropriately. Since the group tokens are generated with the same luminance and random color description, this further breaks down the statistical correlation between luminance and color words.

Finally, we jointly optimize L_δ and L_{ctr} as:

$$L_{\text{total}} = \alpha L_\delta + \beta L_{\text{ctr}}, \tag{16}$$

where we set $\alpha = 1$ and $\beta = -0.0001$.

4. Datasets

To make a fair comparison, we conduct experiments on the extended COCO-Stuff dataset following previous works [6, 42]. Additionally, we collect the multi-instance dataset to facilitate the evaluation of instance-aware colorization.

Extended COCO-Stuff dataset. L-CoDe [42] discards the samples from the COCO-Stuff [3] dataset that does not provide any color description in their captions, leaving 59K training photos and 2.4K validation images. Furthermore, it manually annotates correspondences between color words and object words.

Multi-instance dataset. To provide a large number of samples with distinctive visual characteristics and detailed language descriptions for multiple instances in images, we collect the multi-instance dataset from the related tasks [21, 30]. Our multi-instance dataset includes 65K training images and 7K validation images and each one has a corresponding language description. No correspondence between color words and object words is provided.

5. Experiments

Quantitative evaluation metrics. We separately present Peak Signal-to-Noise Ratio (PSNR) [18], Structural Similarity Index Measure (SSIM) [40], and Learned Perceptual Image Patch Similarity (LPIPS) [50] to evaluate the colorization quality. Moreover, we conduct user studies to determine whether or not human observers consider our results to be favorable.

Training details. We train L-CoIns 80 epochs with batch-size 64 for about 40 hours. We use AdamW optimizer to minimize our losses with learning rate as 1×10^{-5}, momentum parameters $\beta_1 = 0.99$ and $\beta_2 = 0.999$. All experiments are conducted on 8 NVIDIA GeForce RTX 3090 graphic cards.

5.1. Comparison with State-of-the-Art Methods

We make comparisons with a set of language-based colorization algorithms, including LBIE [8], ML2018 [29], Xie2018 [45], L-CoDe [42] and L-CoDer [6]. We retrain and analyze LBIE [8], ML2018 [29] on two separate datasets. Note that we can only train Xie2018 [45], L-Code [42], and L-CoDer [6] on the extended COCO-Stuff dataset and then evaluate them on both datasets for the reason that they require additional parsing mask or additional correspondence between color words and object words as the training guidance.

Qualitative comparisons. We first show visual quality comparisons with the methods above in Fig. 4 with four samples corresponding to different language descriptions, including the description with clear correspondences between color words and object words, the description that assigns colors to instances corresponding to the same object words one by one, the description with unobserved correspondence between color words and object words, and the description against statistical correlation between luminance and color words from top to bottom (a detailed explanation has been provided in Fig. 1). Among these results, our method provides an overall improvement in synthesizing visually pleasing and description-consistent colorization results.

Quantitative comparisons. We present the results of the quantitative comparison in Tab. 1. On the extended COCO-Stuff dataset, our method scores slightly higher than the second place without using additional correspondence annotations, which allows users to describe their intentions more flexibly. On the multi-instance dataset, our method achieves a improvement in achieving instance-aware colorization, demonstrating its superior performance.

5.2. User Study

To evaluate whether our approach could synthesize more appealing colorization results, we conduct two user studies. *(i)* Reality experiment (Reality expt): After being

Figure 4. Qualitative comparison with state-of-the-art methods. **First row:** Our method correctly colorizes all corresponding regions (two glasses). **Second row:** Our method assigns the distinct color to each corresponding instance (left and right men). **Third row:** Our method exactly understands the unobserved correspondence (coral skirts). **Fourth row:** Our method shows robustness for the luminance (purple colorizes the left brighter toy).

Table 2. User study results. Ours (L-CoIns) clearly produces a higher score than previous approaches on both datasets.

Comparison with state-of-the-art methods				
Method	Extended COCO-Stuff		Multi-instance	
	Reality expt	Corresp expt	Reality expt	Corresp expt
LBIE [8]	2.92%	6.88%	3.24%	3.44%
ML2018 [29]	4.20%	5.04%	6.24%	6.48%
Xie2018 [45]	5.24%	9.72%	4.28%	9.96%
L-CoDe [42]	12.88%	18.36%	9.32%	15.04%
L-CoDer [6]	21.32%	28.08%	19.08%	28.56%
Ours	25.40%	**31.92%**	27.64%	**36.52%**
Ground truth	**28.04%**	N/A	**30.20%**	N/A

given a caption that describes the color image, participants are asked to select the image that they believe to be the most visually realistic between the real image and images generated by the six language-based colorization methods. *(ii)* Corresponding experiment (Corresp expt): We apply language-based colorization methods with the specific color description whose color description is replaced randomly to colorize images. Participants were instructed to select the image that closely matches the modified caption.

For each experiment, 100 images from the testing set of our multi-instance dataset are randomly selected. Experiments are separately performed by 25 volunteers and published on Amazon Mechanical Turk (AMT). Our approach achieves highest scores in both studies, as shown in Tab. 2.

5.3. Ablation Study

We disable various modules to create four baselines to study the impact of our proposed modules. The evaluation scores and colorization results of the ablation study are separately shown in Tab. 1 and Fig. 5.

W/o GE (group embedding). We remove the group token along with all related designs, leaving only the counter-color loss calculated between image tokens and language tokens. Without group tokens to aggregate similar image patches, the obvious mismatch between color and instance occurs (first rows in Fig. 5, underside of the car).

W/o GA (grouping attention). We adopt the transformer block in ViT [15] to replace the grouping block. This weakens the ability of the group token to integrate cross-modal information, resulting in an insufficient language understanding and inaccurate colorization results (second row in Fig. 5, yellow cup on the left).

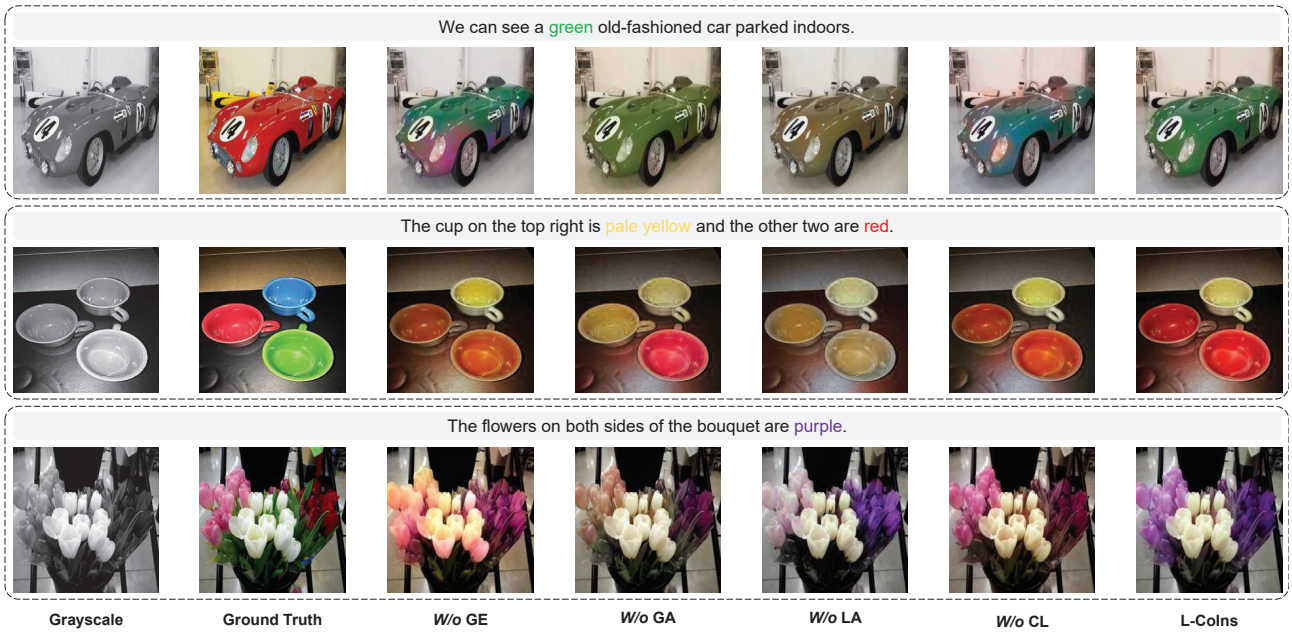

Figure 5. Ablation study with different variants of the proposed method. The colorization results become inconsistent with the language description when some parts of our proposed modules are disabled.

Figure 6. Applying our method to colorize legacy photos.

W/o LA (luminance augmentation). We disable the luminance augmentation in this ablation variant, which allows the model to infer corresponding color words for instance regions based on the statistical correlation between luminance and color words. As a result, it is difficult for the model to colorize special colors for the description is against statistical correlation (second row in Fig. 5, ignoring the word "red").

W/o CL (counter-color loss). We remove the counter-color loss, which reduces the robustness of luminance and eliminates the optimization for the group error. In this way, overall colorization quality downgrades (third row in Fig. 5, incorrectly colorized flowers).

5.4. Application

The application for colorizing legacy black-and-white photos (all are in-the-wild images unseen during training) with distinct language descriptions provided by the user demonstrates our generalization capability in Fig. 6.

6. Conclusion

In this paper, we propose **L**anguage-based **Co**lorization with **Ins**tance awareness (**L-CoIns**). Our model does not require additional external knowledge and has strong robustness for luminance so that the colorization results are more description-consistent. Compared with the state-of-the-art methods, our method achieves the highest PSNR, SSIM, and LPIPS scores on both the existing COCO-Stuff dataset and our collected multi-instance dataset.

Limitation. In the absence of additional annotations (*e.g.*, fine-grained bounding boxes or parsing mask), it remains difficult for our model to capture regions of small objects corresponding to color words in a long caption with detailed descriptions. With the development of unsupervised object detection, this situation could be improved.

Acknowledgements

This work is supported by the National Key R&D Program of China under Grant No. 2021ZD0109803, the National Natural Science Foundation of China under Grand No. 62136001, 62088102, MoE-CMCC "Artifical Intelligence" Project No. MCM20190701, and program for Youth Innovative Research Team of BUPT No. 2023QNTD02.

References

[1] Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Guided image generation with conditional invertible neural networks. *arXiv preprint arXiv:1907.02392*, 2019. 2

[2] Yunpeng Bai, Chao Dong, Zenghao Chai, Andong Wang, Zhengzhuo Xu, and Chun Yuan. Semantic-sparse colorization network for deep exemplar-based colorization. In *ECCV*, 2022. 1

[3] Holger Caesar, Jasper Uijlings, and Vittorio. Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, 2018. 6

[4] Yun Cao, Zhiming Zhou, Weinan Zhang, and Yong Yu. Unsupervised diverse colorization via generative adversarial networks. In *ECML-PKDD*, 2017. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[6] Zheng Chang, Shuchen Weng, Yu Li, Si Li, and Boxin Shi. L-CoDer: Language-based colorization with color-object decoupling transformer. In *ECCV*, 2022. 1, 2, 3, 5, 6, 7

[7] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, 2021. 3

[8] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. Language-based image editing with recurrent attentive models. In *CVPR*, 2018. 2, 6, 7

[9] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *ICCV*, 2015. 2

[10] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-end visual grounding with transformers. In *ICCV*, 2021. 3

[11] Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, and David Forsyth. Learning diverse image colorization. In *CVPR*, 2017. 2

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 4

[13] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021. 3

[14] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. CogView: Mastering text-to-image generation via transformers. In *NIPS*, 2021. 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 3, 4, 7

[16] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. Deep exemplar-based colorization. *TOG*, 2018. 1

[17] Zhitong Huang, Nanxuan Zhao, and Jing Liao. UniColor: A unified framework for multi-modal colorization with transformer. In *SIGGRAPH Asia*, 2022. 2

[18] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*, 2008. 6

[19] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *TOG*, 2016. 2

[20] Xiaozhong Ji, Boyuan Jiang, Donghao Luo, Guangpin Tao, Wenqing Chu, Zhifeng Xie, Chengjie Wang, and Ying Tai. ColorFormer: Image colorization via color memory assisted hybrid-attention transformer. In *ECCV*, 2022. 2, 3

[21] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR - Modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 6

[22] Geonung Kim, Kyoungkook Kang, Seongtae Kim, Hwayoon Lee, Sehoon Kim, Jonghyun Kim, Seung-Hwan Baek, and Sunghyun Cho. BigColor: Colorization using a generative color prior for natural images. In *ECCV*, 2022. 2

[23] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 4

[24] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. In *ICLR*, 2021. 2, 3

[25] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *ECCV*, 2016. 2

[26] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. MAT: Mask-aware transformer for large hole image inpainting. In *CVPR*, 2022. 3

[27] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCV*, 2021. 3

[28] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. FuseFormer: Fusing fine-grained information in transformers for video inpainting. In *ICCV*, 2021. 3

[29] Varun Manjunatha, Mohit Iyyer, Jordan Boyd-Graber, and Larry Davis. Learning to color from language. In *NAACL*, 2018. 1, 2, 6, 7

[30] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 6

[31] Mengxue Qu, Yu Wu, Wu Liu, Qiqi Gong, Xiaodan Liang, Olga Russakovsky, Yao Zhao, and Yunchao Wei. SiRi: A simple selective retraining mechanism for transformer-based visual grounding. In *ECCV*, 2022. 3

[32] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 3

[33] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *CVPR*, 2017. 1

[34] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 3

[35] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *CVPR*, 2020. 2

[36] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NIPS*, 2017. 5

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 3

[38] Patricia Vitoria, Lara Raad, and Coloma Ballester. ChromaGAN: Adversarial picture colorization with semantic class distribution. In *WACV*, 2020. 2

[39] Yi Wang, Menghan Xia, Lu Qi, Jing Shao, and Yu Qiao. PalGAN: Image colorization with palette generative adversarial networks. In *ECCV*, 2022. 2

[40] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *TIP*, 2004. 6

[41] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. CT^2: Colorization transformer via color tokens. In *ECCV*, 2022. 2, 3

[42] Shuchen Weng, Hao Wu, Zheng Chang Chang, Jiajun Tang, Si Li, and Boxin Shi. L-CoDe: Language-based colorization using color-object decoupled conditions. In *AAAI*, 2022. 1, 2, 5, 6, 7

[43] Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. Towards vivid and diverse image colorization with generative color prior. In *ICCV*, 2021. 2

[44] Menghan Xia, Wenbo Hu, Tien-Tsin Wong, and Jue Wang. Disentangled image colorization via global anchors. *TOG*, 2022. 2

[45] Yanping Xie. Language-guided image colorization. Master's thesis, ETH Zurich, Departement of Computer Science, 2018. 2, 6, 7

[46] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *CVPR*, 2022. 3, 5

[47] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. Stylization-based architecture for fast deep exemplar colorization. In *CVPR*, 2020. 1

[48] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. LAVT: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022. 3

[49] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2

[50] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[51] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *TOG*, 2017. 1

[52] Jiaojiao Zhao, Jungong Han, Ling Shao, and Cees GM Snoek. Pixelated semantic colorization. *IJCV*, 2020. 2

[53] Jiaojiao Zhao, Li Liu, Cees GM Snoek, Jungong Han, and Ling Shao. Pixel-level semantics guided image colorization. In *BMVC*, 2018. 2

[54] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 3

[55] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 3

[56] Changqing Zou, Haoran Mo, Chengying Gao, Ruofei Du, and Hongbo Fu. Language-based colorization of scene sketches. *TOG*, 2019. 2