Automatic Controllable Colorization via Imagination

Xiaoyan Cong[3]   Yue Wu[4]   Qifeng Chen[2]   Chenyang Lei[1*]

[1]CAIR, HKISI-CAS   [2]HKUST   [3]Zhejiang University   [4]Huawei Noah's Ark Lab

Abstract

We propose a framework for automatic colorization that allows for iterative editing and modifications. The core of our framework lies in an imagination module: by understanding the content within a grayscale image, we utilize a pre-trained image generation model to generate multiple images that contain the same content. These images serve as references for coloring, mimicking the process of human experts. As the synthesized images can be imperfect or different from the original grayscale image, we propose a Reference Refinement Module to select the optimal reference composition. Unlike most previous end-to-end automatic colorization algorithms, our framework allows for iterative and localized modifications of the colorization results because we explicitly model the coloring samples. Extensive experiments demonstrate the superiority of our framework over existing automatic colorization algorithms in editability and flexibility. Project page: https://xy-cong.github.io/imagine-colorization/.

1. Introduction

Countless black-and-white photographs, relics of bygone eras, exist today, often requiring the deft touch of skilled artists to breathe color into them. However, such a manual process is labor-intensive and time-consuming. Given a grayscale image, colorization methods aim to estimate the corresponding color automatically or with other guidance. Achieving expert-level colorization is challenging due to the multi-modality nature of the task: there exist multiple plausible colorization results for one grayscale image.

Automatic colorization algorithms [27, 35, 64, 72, 74, 75, 85, 86] have made significant progress in recent years and have demonstrated impressive results in many common scenarios. However, they are still hard to match the colorization proficiency of human experts. Many data-driven methods struggle with issues such as grayish and desaturation because of the color ambiguity that one object may present multiple plausible colors in datasets. GAN-based approaches tend to produce unpleasant artifacts due to the lim-

Figure 1. **Our colorization results.** By leveraging our designed Imagination Module, our framework can achieve photorealistic and vivid colorization results.

ited representation capacity in GAN's prior. DDColor [32] alleviates these shortcomings to some extent but is unable to achieve controllable and diverse colorization. Human experts excel in colorizing black-and-white photos due to their rich and imaginative priors, enabling them to transform grayscale images into visually pleasant renditions. They possess instance-aware and semantic-aware understanding both locally and globally, imagining diverse plausible colors and assigning the most appropriate ones based on context. Exemplar-based colorization methods [3, 11, 47, 80] achieve pleasing results through emulating the human experts' priors, wherein a highly-corresponding reference is manually provided. However, identifying a perfect reference can be both labor-intensive and detract from an optimal user experience.

Another challenge with most automatic algorithms is that the obtained results are often difficult to modify and improve. For example, Su et al. [64] and Zhang et al. [85] can only produce one deterministic result. While UniColor [27] offers an interactive colorization, it demands users to provide additional information, such as text or strokes.

In this paper, we propose a novel automatic colorization framework that allows iterative editing and modifications from users. Our framework is designed to mimic the process of human expert imagination colorization, as shown in Figure 2. We introduce an imagination module designed to provide semantically similar, structurally aligned, and instance-aware references, mirroring human imagination. Our framework begins by composing a refined reference from multiple candidates generated by pretrained cross-modality image generation models, such as ControlNet [82]. The rich color priors embedded in large pretrained Diffusion models can be harnessed to synthesize references with diverse and col-
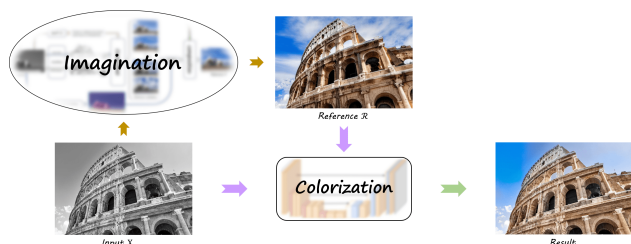
---
*Corresponding author

Figure 2. **Framework Overview.** Given a black-and-white input, our framework first synthesizes a semantically similar, spatially aligned, and instance-aware reference by emulating the imagination process of human experts. Then the colorization module colorizes the black-and-white image with the guidance of reference. On the right, we present an additional example.

orful colors. Then, our colorization module colorizes the black-and-white input under the guidance of the reference.

Our composition strategy involves two steps: First, we carry out the semantic segmentation of the input. Second, we compose the refined reference by selecting the segment most similar to all potential reference candidates. The composition strategy enables our framework to exhibit diverse colorization results, as well as provides the flexibility for controllable and editable colorization tailored to the unique desires and preferences of the user. Specifically, if users find certain segments unsatisfactory, they can effortlessly substitute those segments with alternatives from the pool of reference candidates or use an image of their choosing. Consequently, the colorization results are accordingly modified to reflect these changes.

Our contributions can be summarized as follows:

- We propose a novel automatic colorization framework that leverages the pre-trained diffusion models. We introduce an imagination module that emulates human experts to synthesize semantically similar, structurally aligned, and instance-aware colorful references, with potential applications beyond colorization.
- We demonstrate our novel automatic colorization framework exhibits remarkable controllable and user-interactive capabilities. We can also present diverse colorization results.
- Compared to previous automatic colorization methods, our framework achieves state-of-the-art performance and generalization.

2. Related Work

Automatic Colorization. Deep learning based methods [1, 18, 27, 28, 32, 39–41, 64, 69, 72, 74, 75, 81, 85, 89] achieve automatic colorization by learning the relationship between semantic and color on large datasets. Additional semantic information is incorporated to improve colorfulness, such as classification [28, 40, 85, 86], semantic segmentation [18, 89] and instance awareness [64]. The performance of the above methods is usually good when the color of the object (e.g., a green tree) is closely related to the semantic, while they tend to produce muted or grayish results for objects with a wide spectrum of potential colors (e.g., T-shirt). In contrast, our framework can alleviate the limitation and generate more diverse and colorful results thanks to our innovative imagination module.

User-guided Colorization. Most classical approaches rely on user guidance for colorization, including scribble-based colorization, text-based colorization, and example-based colorization. Scribble-based methods [42, 50, 55, 60, 83, 84] usually formulate colorization as an optimization problem with some assumptions. For example, Levin et al. [42] assume pixels with similar intensity share similar colors. Some follow-up methods work on achieving more efficient color propagation by leveraging texture similarity [50, 55] and intrinsic distance [79]. These approaches often require highly labor-intensive manual processes. Text-based methods [2, 7–9, 52, 73, 77] offer a reprieve from this obstacle by leveraging user-friendly language descriptions delineating objects and their corresponding colors to guild colorization. Some recent approaches aim to enhance the robustness and consistency of colorization by the grouping mechanism [9] aggregating similar image patches and assigning vivid colors for unmentioned instances [8].

Example-based methods [3, 4, 10–12, 12, 16, 19, 22, 26, 29, 44, 47–49, 68, 71, 78, 80, 81, 88] further reduce the difficulty of colorization for normal users. Only one reference image is required for automatic colorization. Some approaches [3, 26, 78, 88] focus on leveraging AdaIN [25] to transfer the global color distribution of reference image, while others [3, 4, 11, 12, 22, 29, 44, 68, 81] transfer color according to pixel-level semantic correspondence. Although these methods have improved colorization performance a lot, it is still non-trivial, time-consuming, and not user-friendly for users to find a perfect reference by searching on the Internet or providing descriptions and segmentation cues to an image retrieval system [12]. In our work, the imagination module avoids such efforts by automatically synthesizing a semantically similar and structurally aligned colorful reference image.

Generative Priors Colorization. Generative priors embedded in pretrained GANs and Diffusions have been pivotal in achieving photorealistic colorization. GANs [17, 33, 34, 46, 54, 65–67, 70] are a huge family of generative models for conditional and unconditional image synthesis. Some specific conditions are utilized to control better image synthesis, such as layout [65, 66] and semantic map [46, 54, 67, 70] et al. StyleGAN-family [33, 34] models have shown impressive performance in unconditional high-resolution image synthesis. Diffusion probabilistic models [14, 24, 36, 38, 56, 57, 59, 62, 63, 82] have emerged as another formidable family. Thanks to the Denoising Diffusion Probabilistic Model (DDPM) [24] and the Denoising Diffusion Implicit Model (DDIM) [63], Latent Diffusion
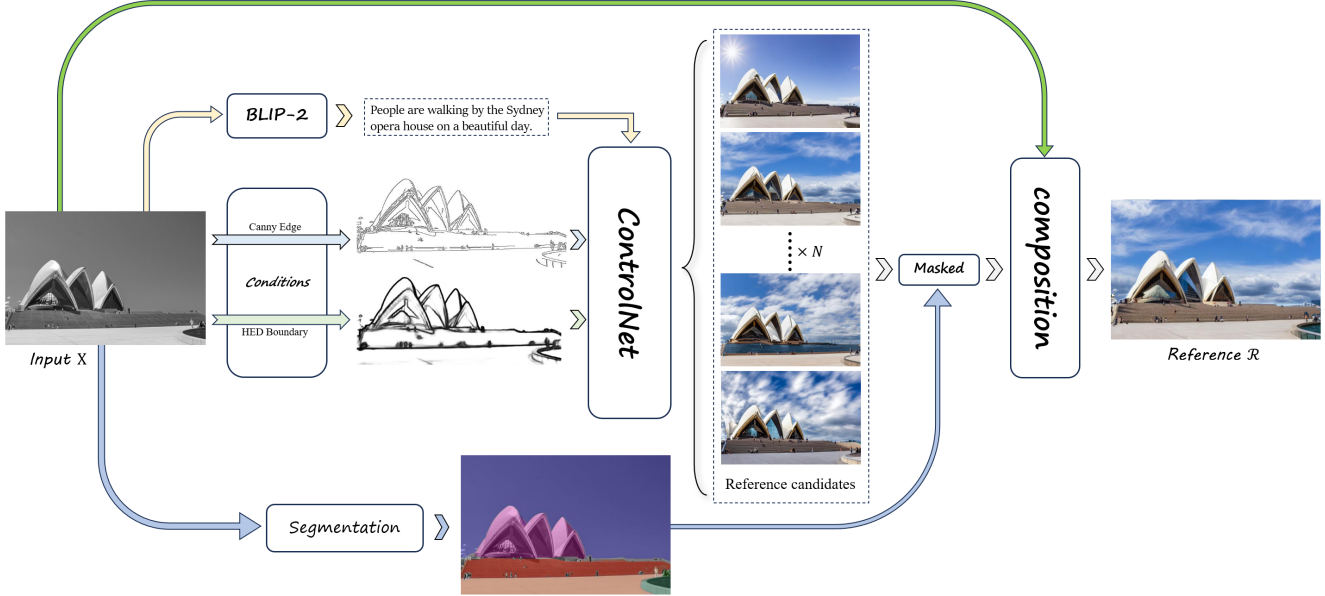
Figure 3. **Imagination Module and Reference Refinement Module.** In Imagination Module 3.3, given a grayscale input, we generate N reference candidates $\mathbf{C}$, N_1 of which are conditioned on the canny edge of $\mathbf{X}$, and N_2 of which are conditioned on the HED boundary of $\mathbf{X}$, $N = N_1 + N_2$. In Reference Refinement Module 3.4, we first extract the segmentation $\mathbf{S}$ of $\mathbf{X}$. For each segment $\mathbf{S}^j$, we look for the best reference segment $\mathbf{R} \odot \mathbf{S}^j$ for the optimal reference $\mathbf{R}$ by selecting the nearest neighbor for $\mathbf{X} \odot \mathbf{S}^j$ among all reference candidates $\mathbf{C}_i \odot \mathbf{S}^j$ in terms of the difference in the robust and universe DINOv2 [53] feature space.
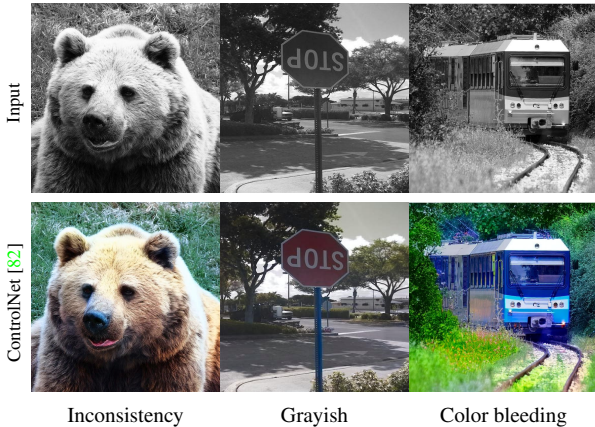

Figure 4. **ControlNet Issues.** Directly applying ControlNet[82] to colorization might cause issues like inconsistency, grayish, and color bleeding. We use gray images as conditions of ControlNet and corresponding color images as ground truth. A pretrained stabel diffusion model is used as the initialization. We then train the model on the commonly used colorization datasets.

Model (LDM) like Stable Diffusion [56] achieves state-of-the-art performance in text-to-image generation tasks. Recognizing the vast potential of LDM, Controlnet [82] controls pretrained diffusion models with task-specific conditions, accommodating multi-modal inputs, and thereby expanding a wider range of application scenarios. While these models struggle to maintain the local spatial structures of grayscale inputs, their inherent capability to harness the diverse color priors from pretrained models offers a promising avenue for synthesizing reference images.

3. Method

3.1. Overview

Given a grayscale image $\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$ as input, our framework automatically predicts its color channels $\mathbf{Y} \in \mathbb{R}^{H \times W \times 2}$. $\mathbf{X}$ and $\mathbf{Y}$ are the lightness and ab channel in the CIE Lab color space. H and W are the height and width of the grayscale image. The key difference between our approach and existing automatic colorization methods is that: we leverage diverse color priors embedded in large pretrained generative models, such as Stable Diffusion, to synthesize colorful images based on the semantic context and spatial structure of the input grayscale image as a reference for colorization.

Figure 2 shows the architecture of our framework. We first utilize ControlNet[82] to control a large diffusion model to synthesize semantically similar and structurally aligned reference candidates $\mathbf{C} \in \mathbb{R}^{N \times H \times W \times 3}$ based on the grayscale input $\mathbf{X}$ from our imagination module $\mathcal{I}$:

$$\mathbf{C} = \mathcal{I}(\mathbf{X}, \mathbf{c}(\mathbf{X})), \tag{1}$$

where $\mathbf{c}(\mathbf{X})$ means various optional conditions based on $\mathbf{X}$. The reference $\mathbf{R} \in \mathbb{R}^{H \times W \times 3}$ is synthesized by a reference refinement module $\mathcal{P}$ from multiple reference candidates $\mathbf{C}$. Then, our example-based colorization module $\mathcal{C}$ takes the grayscale image $\mathbf{X}$ and synthesized reference $\mathbf{R}$ as input,

Figure 5. **Diverse Colorization.** We can synthesize diverse colorful references from Imagination Module 3.3, yielding diverse colorzation results.

and outputs the corresponding photorealistic colorization results:

$$(\mathbf{X}, \mathbf{Y}) = \mathcal{C}(\mathbf{X}, \mathbf{R}), \tag{2}$$

$$\mathbf{R} = \mathcal{P}(\mathbf{C}). \tag{3}$$

The details of the imagination module, reference refinement module and colorization module are illustrated in Section 3.3, Section 3.4 and Section 3.5.

3.2. Preliminary

Image Diffusion Models. Image Diffusion Models address the image generation task by learning to perform progressive denoising [14, 24, 56, 58]. Given an image $\mathbf{I}_0$, noise is progressively added to produce $\mathbf{I}_t$, a noisy version of $\mathbf{I}_0$. Along with t increasing, the image is gradually approaching pure noise. Image diffusion models learn a neural network ϵ_θ to predict the noise added to the noisy image $\mathbf{I}_t$ and the objective function can be simplified as

$$\mathcal{L} = \mathbb{E}_{\mathbf{I}_0, t, \mathbf{c}, \epsilon \in \mathcal{N}(0,1)}[\|\epsilon - \epsilon_\theta(\mathbf{I}_t, t, \mathbf{c})\|_2^2]. \tag{4}$$

Successful image diffusion models like Stable Diffusion [56] operate in the latent space of an autoencoder, making convergency more stable and faster.

3.3. Imagination Module

Given a grayscale image $\mathbf{X}$, our imagination module aims to synthesize multiple semantically similar and structurally aligned reference candidates $\mathbf{C}$. The motivation is to imitate human imagination: humans can achieve photorealistic colorization because they possess sufficient and realistic color priors regarding our colorful world. For each part (e.g. instance) within the grayscale image, we inherently harbor diverse and colorful examples, which empowers us to perform a locally and globally consistent photorealistic composition by selectively choosing from various examples for each part. Similarly, the reference candidates synthesized by our imagination module should encapsulate such color diversity and generalization attributes. Thus, pre-trained large image diffusion models such as Stable Diffusion are ideally suited as our synthesis model, as they inherently contain diverse color priors.

In this work, we utilize ControlNet [82] to control large pretrained image diffusion models to synthesize a total of N
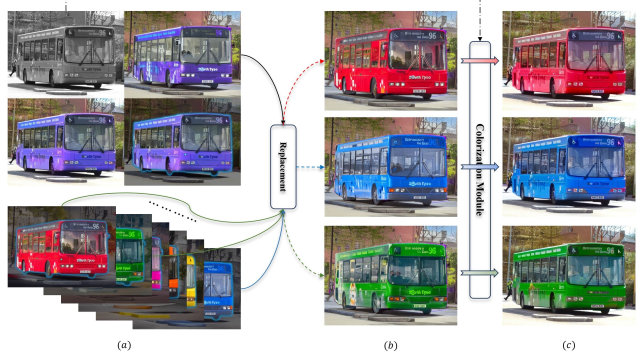

Figure 6. **User Interactive.** The top panel of (a) shows the black-and-white input, the reference, the colorization result, and the area marked by the user's mouse click indicating dissatisfaction with the colorization. The bottom panel of (a) displays multiple reference candidates generated by our Imagination Module. (b) Users select a preferred segment from the reference candidates to replace the unsatisfactory part of the reference, consisting of a new reference. (c) The colorization result after the adjustment.

reference candidates given colorization-task-specific input conditions, with the help of captions generated by BLIP-2 [45]. Numerous conditional inputs are available, including canny edge [6], HED boundary [76], semantic segmentation map and so on. The overarching objective, irrespective of the chosen conditional input methods, is to furnish a diverse, colorful, and photorealistic color space. We provide an empirical approach: Given the grayscale input $\mathbf{X}$, we generate a total of N reference candidates $\mathbf{C}$, N_1 of which are conditioned on the canny edge of $\mathbf{X}$, and N_2 of which are conditioned on the HED boundary of $\mathbf{X}$, $N = N_1 + N_2$. Notably, the direct application of ControlNet to colorization leads to issues like inconsistency, grayish, and color bleeding, as shown in Figure 4. In our imagination module, however, ControlNet shows extremely superior performance on synthesizing diverse and colorful reference candidates, providing quite strong priors for the following colorization module.

3.4. Reference Refinement Module

Given N reference candidates $\mathbf{C} \in \mathbb{R}^{N \times H \times W \times 3}$ generated from our imagination module and the grayscale input $\mathbf{X}$, our reference refinement module $\mathcal{P}$ aims to synthesize the most semantically similar and structurally aligned refined reference $\mathbf{R} \in \mathbb{R}^{\times H \times W \times 3}$ through aggregating information across the diverse and colorful reference candidates space. Our motivation is that although a single reference candidate $\mathbf{R} = \mathcal{C}_i$ is enough to provide required priors and guidance in many cases, it is still worth performance improvement further by aggregating diverse information in multiple reference candidates to solve the problem: while $\mathcal{C}_i$ is semantically similar and structurally aligned with $\mathbf{X}$, there might be some incompatible mismatched segmentations in terms of luminance. For example, a desk with high luminance tends to be

a bright color, but the generative model may offer a dark one. Both of them share the same grayscale input and ControlNet conditions like canny edge and HED boundary images.

We address this problem by composing a refined reference $\mathbf{R}$ from all multiple reference candidates $\mathbf{C}$. Specifically, we first use an off-the-shelf segmentation model [20, 37, 43, 90] to obtain the segmentation of $\mathbf{X}$. In our implementation, we employ the Semantic-SAM [43] to generate semantic-aware and instance-aware segments $\mathbf{S} = \text{Semantic-SAM}(\mathbf{X}) = \{\mathbf{S}^j, j = 1, 2, \ldots, N_\mathbf{S}\}$, where we denote each segment as $\mathbf{S}^j$. As $N_\mathbf{S}$ decreases, the composed reference becomes more inclined to capture the global information embedded in the reference candidates. Conversely, as $N_\mathbf{S}$ increases, the composed reference tends to capture more subtle details among the reference candidates. For each segment $\mathbf{S}^j$, we look for the best reference segment $\mathbf{R} \odot \mathbf{S}^j$ by selecting the nearest neighbor for $\mathbf{X} \odot \mathbf{S}^j$ among all reference candidates $\{\mathbf{C}_i \odot \mathbf{S}^j\}_{i=0}^N$ in terms of the difference in the universe feature space:

$$\alpha(j) = \underset{i}{\arg\min} \, \mathcal{M}(\mathcal{D}(\mathbf{X} \odot \mathbf{S}^j), \mathcal{D}(\mathbf{C}_i \odot \mathbf{S}^j)), \quad (5)$$

where $\odot$ denotes operation to select some regions of an image, $\alpha(j)$ denotes the best reference candidate index for segment $\mathbf{S}^j$, $\mathcal{M}$ denotes distance measurement metrics ($\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_{\text{cosine}}$, etc.), and $\mathcal{D}(\cdot)$ denotes the robust and universe image feature DINOv2 [53]. Then, our refined semantically similar and structurally aligned reference R can be composed as

$$\mathbf{R} = \bigcup_{j=0}^{N_\mathbf{S}} (S^j \odot C_{\alpha(j)}). \quad (6)$$

User-interactive refinement. So far, a semantically similar and structurally aligned reference can be synthesized automatically. Although such a reference is capable of achieving visually pleasing colorization, the color of each segment may not be the color that users are most satisfied with. Here we provide a user-interactive reference adjustment method: When users are unsatisfied about some segments $\mathbf{S}^j$, they can effortlessly substitute those segments $\mathbf{R} \odot \mathbf{S}^j$ with alternatives from the pool of reference candidates $\{\mathbf{C}_i \odot \mathbf{S}^j\}_{i=0}^N$ or use an image of their choosing. For example, in Figure 6, we replace the bus with buses with different colors and preserve the colors in other areas. The colorization results are accordingly adjusted. In this case, we achieve controllable and editable colorization, which is non-trivial in the automatic colorization field.

3.5. Colorization Module

Our colorization module draws inspiration from the unified colorization framework UniColor [27] to colorize the grayscale input $\mathbf{X} \in \mathbb{R}^{H \times W}$ by utilizing the reference image $\mathbf{R} \in \mathbb{R}^{H \times W \times 3}$ synthesized by our imagination module.

Input　　*Coarse Hint Colors*　　*Coarse Results*

Reference　　*Fine Hint Colors*　　*Fine Results*

Figure 7. **Coarse-to-fine hint colors optimization** can mitigate the color ambiguity within semantic instances by identifying the most representative hint colors, as highlighted by the boxed area in the figure.

UniColor is a unified colorization framework based on multi-modal conditions. Multi-modal conditions, such as strokes, reference images, and text, are converted into hint colors which are leveraged as a guidance for subsequent colorization. Specifically, the colorization process is divided into two steps: coarse-to-fine hint colors generation and hint colors propagation.

Coarse-to-fine hint colors generation. Hint color points are generated according to the semantic correspondence between $\mathbf{X}$ and $\mathbf{R}$ as an intermediate representation to instruct the following colorization process. Initially, we employ the correspondence network proposed in [81] to compute the pairwise similarity between the features of $\mathbf{X}$ and the features of $\mathbf{R}$ extracted by VGG19 [61], serving as the semantic correspondence, through which the color of $\mathbf{R}$ can be warped onto $\mathbf{X}$, resulting in a warped image $\mathbf{W}$. Similar to UniColor [27], $\mathbf{X}$ is divided into cells with size $d \times d$. If a cell's average semantic similarity surpasses the threshold value $\mathbf{s}_\epsilon$, the mean color of that cell in the warped image is designated as a hint color. This helps in preventing the mismatching area as the hint colors. However, we observe that merely employing the aforementioned method to generate hint colors can lead to semantic color ambiguity, as illustrated in Figure 7. This is due to a distinction from the conventional reference-based colorization approach where a semantically similar, high-quality reference is manually provided. In our imagination module, the process of synthesizing $\mathbf{R}$ might introduce color inconsistency within semantic masks. Therefore, we treat the hint colors derived above as a coarse hint colors set $\mathcal{H}_{coarse}$ and obtain the fine hint colors set $\mathcal{H}_{fine}$ through optimization. Utilizing the Semantic-SAM segments $\mathbf{S} = \{\mathbf{S}^j, j = 1, 2, \ldots, N_S\}$ from the Reference Refinement Module, we associate each segment $\mathbf{S}^j$ with the corresponding coarse hint colors in $\mathcal{H}_{coarse}$. If the number of hint colors in $\mathbf{S}^j$ exceeds $\mathcal{N}_\mathcal{H}$, we cluster the hint colors in $\mathbf{S}^j$ to identify and eliminate any hint colors with anomalous color values from the color distribution of most hint colors. In our implementation, we derive the fine hint colors set by leveraging

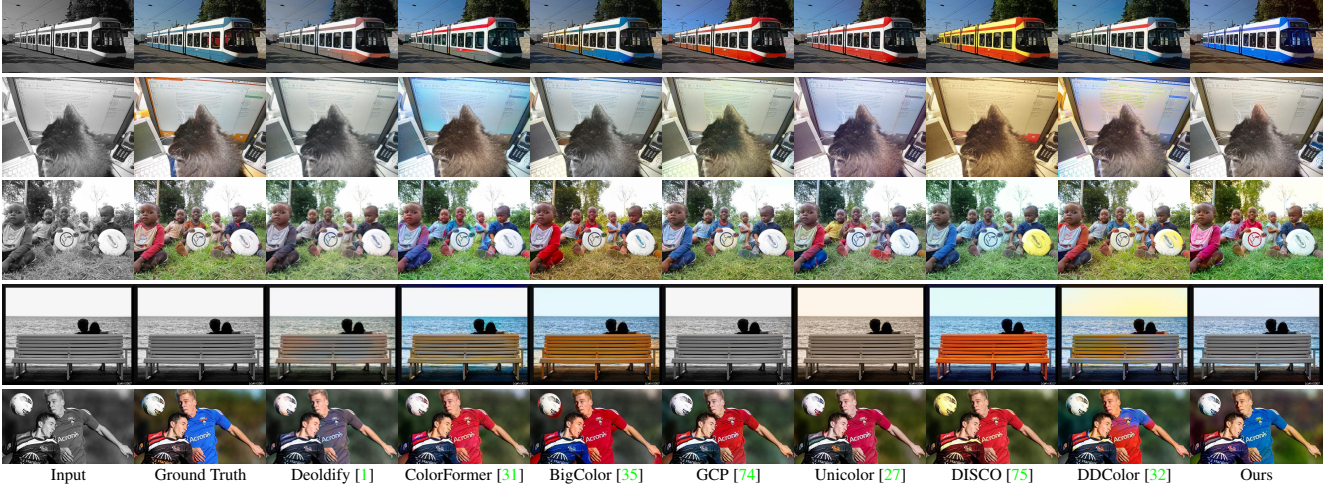| Input | Ground Truth | Deoldify [1] | ColorFormer [31] | BigColor [35] | GCP [74] | Unicolor [27] | DISCO [75] | DDColor [32] | Ours |

Figure 8. **Qualitative comparisons on the COCO validation set.** Our method can generate more natural and photo-realistic colors than state-of-the-art approaches. Zoom in for the best view.

the DBSCAN algorithm [15, 30] for clustering.

Hint colors propagation. Similar to UniColor [27], given a black-and-white input and fine hint colors, a Hybrid-Transformer is harnessed to autoregressively predict chrominance features, which are then concatenated with luminance features extracted from the input. The concatenated features are fed into a decoder to produce the final colorization result. We refer readers to UniColor [27] for more details.

4. Experiments

4.1. Experimental setting

Dataset. Our experiments are mainly conducted on standard evaluation benchmark, *i.e.*, COCO-stuff [5] and ImageNet [13]. Both our framework and baselines are evaluated in (i) COCO-stuff validation set (5k images) (ii) ImageNet testing split ctest (10k images), (iii) *In-the-wild* photos collected on the Internet (500 images), qualitatively and quantitatively. Notably, evaluating the colorization performance on arbitrary grayscale images from the Internet, often referred to as *In-the-wild*, is incredibly crucial. This is because they might have a completely different color distribution compared to widely used datasets such as ImageNet and COCO-stuff. Such differences pose a significant challenge to the generalization capability of colorization models. In this regard, our framework significantly outperforms the state-of-the-art baselines.

Evaluation Metrics. In experiments, we mainly care about the framework's performance on perceptual realism and color vividness. For measuring perceptual realism, we adopt Fréchet Inception Distance (**FID**) [23] to compare the distribution similarity between the colorization results and the ground truth. For color vividness, we employ the Colorfulness (**CF**) metric [21] which is close to human vision perception. In addition, following previous works [64, 72, 86],

we provide evaluations on PSNR, SSIM and LIPIS [87] for reference.

Baselines. We adopt the most recent and related automatic colorization methods as baselines, including Deoldify [1], Su et al. [64], Coltran [39], GCP [74], CT2 [72], ColorFormer [31], BigColor [35], UniColor (Automatic) [27], DISCO [75] and DDColor [32]. In our experiments, we directly use the released evaluation results by Su et al [64] and DISCO [75], and we use automatic colorization results by Unicolor [27] without user hints. For the other baselines, we compare using their official codes and weights.

Condition Setting. We provide an ablation study in Table 1 on the condition setup for ControlNet [82]. Various feature maps can be used as conditions for ControlNet, with Canny and HED maps demonstrating superior performance. Notably, the combination of Canny and HED map conditions yields optimal results, prompting our choice of this configuration.

Condition setting	Canny	HED	Depth	Semantic	Canny&HED
CF ↑	46.99	45.28	42.81	45.02	**47.18**

Table 1. **Ablation study on the condition setup.** We generate $N = 8$ reference candidates for each condition setting. The test images are from the COCO-stuff validation dataset.

Hyper-parameters Setup. We provide an ablation study in Table 2 on the hyper-parameter choice: the number of reference candidates N, the number of segments $N_\mathbf{S}$, and the threshold of hint colors $\mathcal{N}_\mathcal{H}$. We empirically select the hyper-parameters as $N = 8$, $N_\mathbf{S} = 10$, and $\mathcal{N}_\mathcal{H} = 10$ based on the balance of performance and running efficiency.

4.2. Comparisons to Baselines

Qualitative Comparisons. We show the qualitative comparisons in Figure 8 and Figure 9. Note that the test images are from COCO-stuff validation set and *In-the-wild* set respectively. Our colorization results are more natural, pho-

Figure 9. **Qualitative comparisons on in the wild validation set.** Our method can generate significantly more natural and photo-realistic colors than state-of-the-art approaches. Zoom in for the best view.

N	2	4	8	16	**32**	8	8
N_S	10	10	10	10	**10**	40	10
$\mathcal{N}_{\mathcal{H}}$	10	10	10	10	**10**	10	50
CF ↑	38.81	43.13	47.18	47.82	**48.69**	42.09	45.53

Table 2. **Ablation study on the hyper-parameters choice: N, N_S and $\mathcal{N}_{\mathcal{H}}$.** The test images are from the COCO-stuff validation dataset.

torealistic, vivid, and visually pleasing than baselines in both scenarios, especially on the *In-the-wild* set, significantly alleviating artifacts and preserving color consistency. Since there might be quite a lot of *unseen* color distributions in *In-the-wild* set, state-of-the-art baselines suffer from color bleeding, grayish colorization, color inconsistency, and visual artifacts. In contrast, our imagination module $\mathcal{I}$ can provide a wide range of reference candidates, exhibiting sufficient color diversity and strong colorization robustness. Our framework has taken a significant leap forward, to some extent, achieving a universal off-the-shelf colorization application. More comparisons are available in the supplementary.

Quantitative Comparisons. We show quantitative comparisons in Table 3. Considering the color multimodality in the colorization task, the ground truth based evaluation metrics PSNR, SSIM, and LPIPS can not accurately reflect the actual performance, for there might exist huge differences between the colorful and plausible colorization results and the ground truth. Our framework shows the best quantitative results in **CF**, indicating that we can achieve the most colorful and perceptually photorealistic colorization results.

User study. We conduct a user study to evaluate which approach is more preferred by human observers. Specifically, we compare our approach with five strong baselines: DDColor [32], DISCO [75], UniColor (Automatic) [27], Bigcolor [35], CT2 [72]. 30 Users are presented with five colorized images (our approach and one from each of the four baselines) simultaneously in a single view, with the positions randomized for each comparison. They are then tasked with selecting the most realistic and colorful one among them. We randomly select 30 images from the COCO-stuff
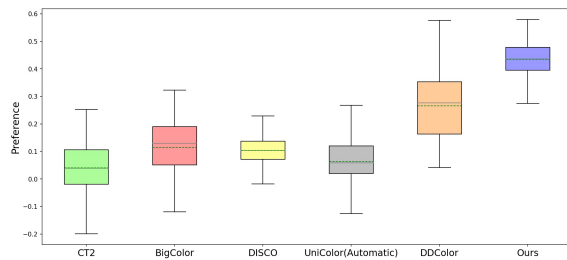

Figure 10. **User study.** The dashed green lines are the mean preference percentage.

validation set, 30 images from the ImageNet validation set, and 20 images sourced from the internet. As shown in Figure 10, Our approach outperforms DDColor [32], DISCO [75], UniColor (Automatic) [27], Bigcolor [35], CT2 [72] with a preferred rate of $43.3\%, 26.7\%, 10.0\%, 6.3\%, 10.0\%, 3.3\%$.

4.3. Ablation Study

Reference refinement module. Figure 11 illustrates the effectiveness of our Reference Refinement Module (see Section 3.4), which can significantly improve the colorfulness and photorealism of colorization results. In the top row, reference candidates are composed to an optimal reference, with '×N' indicating there are N candidates in all. In the bottom row, we showcase colorization results guided by two randomly selected reference candidates and the optimal reference. Upon comparison, Without our composition strategy, the diversity and completeness of the color space are compromised. ControlNet [82] struggles to generate a comprehensive color space from a single reference. Consequently, eliminating the composition strategy compromises the completeness of this color space, leading to inconsistencies, visual artifacts, and color bleeding. Quantitatively, Table 4 further highlights a marked performance difference when using our Reference Refinement Module 3.4 compared to not using it.

Diverse Colorization. Thanks to our imagination module 3.3, we can synthesize multiple reasonable and colorful references, yielding diverse colorization results, as shown in

Method	COCO-Stuff					ImageNet					In-the-wild
	FID ↓	SSIM ↑	PSNR ↑	LPIPS ↓	**CF** ↑	FID ↓	SSIM ↑	PSNR ↑	LPIPS ↓	**CF** ↑	**CF** ↑
Coltran [39]	13.1	0.359	13.3	0.513	38.5	15.5	0.264	8.65	0.723	**55.9**	38.0
GCP [74]	7.35	0.900	22.4	0.194	37.9	3.36	0.931	23.5	0.238	33.9	37.0
CT2 [72]	22.9	0.358	13.5	0.518	46.2	8.44	0.354	14.5	0.480	41.6	44.5
ColorFormer [31]	8.69	0.756	21.3	0.216	41.0	3.83	0.834	22.5	0.196	36.0	39.9
BigColor [35]	8.53	0.832	20.8	0.217	43.4	3.59	0.893	21.5	0.212	40.4	37.8
UniColor [27]	7.90	0.855	22.4	0.195	36.4	6.22	0.909	22.0	0.238	35.7	33.9
DISCO [75]	11.2	0.738	19.4	0.236	46.2	9.21	0.783	21.0	0.265	44.5	43.5
DDColor [32]	6.19	0.904	23.1	0.169	44.7	3.16	0.885	23.4	0.186	42.2	41.9
Ours	7.21	0.859	23.3	0.180	**47.2**	3.62	0.884	23.8	0.207	<u>48.8</u>	**48.9**

Table 3. **Quantitative Comparison with baselines on three datasets.** We provide evaluations on PSNR, SSIM, and LPIPS just for reference because those metrics can not accurately reflect the actual performance. Our framework shows the best quantitative results in **CF**, indicating we can generate the most colorful and perceptually photorealistic results.

Figure 11. **Ablation Study: Reference Refinement Module 3.4.** Without our composition strategy, the diversity and completeness of the color space are compromised. Consequently, the reference exhibits inconsistency, visual artifacts, and color bleeding, leading to unsatisfactory colorization results.

Figure 5.

Hint Colors Optimization. shows that the hint colors ambiguity issue can arise in the absence of a coarse-to-fine optimization strategy. As depicted in Figure 7, the reference's sky exhibits a substantial variation in hint color values, resulting in color bleeding and incorrect colorization. By employing a clustering algorithm, our coarse-to-fine optimization identifies the most representative set of hint colors, which reduces color bleeding and ensures a pleasing colorization result.

4.4. Limitation

As depicted in Figure 12, our approach encounters challenges when dealing with images populated with numerous identical instances. We attribute such cases to the inherent limitations of existing generative models like ControlNet which struggle with generating diverse and photo-realistic images with numerous distinct humans. Thus, we believe our method will address such cases with more advanced generative models emerging. Besides, further enhancement may require the colorization framework to achieve an equilibrium: ensure a consistent color prior for a whole semantic area, while still recognizing and respecting the unique color cues of each instance within. Moreover, the inference cost is another concern. After accelerating ControlNet with LCM-LoRA [51], our model takes about eight seconds to colorize

Input   Ours   Ground truth

Figure 12. **Failure Cases.** Our approach might produce artifacts and inconsistencies when colorizing images with too many identical instances.

reference	FID ↓	CF ↑	SSIM ↑	PSNR ↑	LPIPS ↓
composition	**7.21**	**47.18**	**0.859**	**23.3**	**0.180**
single	8.28	43.87	0.848	21.2	0.230

Table 4. **Ablation on without and with composition strategy.** All the tested images are from COCO-Stuff validation set. Our composition strategy in the Reference Refinement Module plays a pivotal role in improving colorization results from quantitative perspectives.

a 512×512 image on a single Nvidia GeForce RTX 3090.

5. Conclusion

In this paper, we present a novel framework for automatic image colorization that emulates the imagination process of human experts. Our primary contribution is the Imagination Module and the Reference Refinement Module, which together generate reference images that are semantically similar, structurally aligned, and instance-aware. The Colorization Module then colorizes the black-and-white inputs with the guidance of these reference images. Our framework surpasses state-of-the-art baselines both qualitatively and quantitatively, achieving controllable, editable, and diverse colorization results, which is non-trivial in the automatic colorization community. We posit that the concept of imagination has the potential to enhance a variety of computer vision tasks.

References

[1] Jason Antic. jantic/deoldify: A deep learning based project for colorizing and restoring old images (and video!).

[2] Hyojin Bahng, Seungjoo Yoo, Wonwoong Cho, David K. Park, Ziming Wu, Xiaojuan Ma, and Jaegul Choo. Coloring with words: Guiding image colorization through text-based palette generation, 2018.

[3] Yunpeng Bai, Chao Dong, Zenghao Chai, Andong Wang, Zhengzhuo Xu, and Chun Yuan. Semantic-sparse colorization network for deep exemplar-based colorization. In *European Conference on Computer Vision*, pages 505–521. Springer, 2022.

[4] Aurélie Bugeau, Vinh-Thong Ta, and Nicolas Papadakis. Variational exemplar-based image colorization. *IEEE Transactions on Image Processing*, 23(1):298–307, 2013.

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context, 2018.

[6] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

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

[8] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. L-cad: Language-based colorization with any-level descriptions, 2023.

[9] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. L-coins: Language-based colorization with instance awareness. In *CVPR*, 2023.

[10] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf. Automatic image colorization via multimodal predictions. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III 10*, pages 126–139. Springer, 2008.

[11] Siqi Chen, Xueming Li, Xianlin Zhang, Mingdao Wang, Yu Zhang, and Yue Zhang. Spcolor: Semantic prior guided exemplar-based image colorization, 2023.

[12] Alex Yong-Sang Chia, Shaojie Zhuo, Raj Kumar Gupta, Yu-Wing Tai, Siu-Yeung Cho, Ping Tan, and Stephen Lin. Semantic colorization with internet images. *ACM Transactions on Graphics (ToG)*, 30(6):1–8, 2011.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[15] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996.

[16] Faming Fang, Tingting Wang, Tieyong Zeng, and Guixu Zhang. A superpixel-based variational model for image colorization. *IEEE Transactions on Visualization and Computer Graphics*, 26(10):2931–2943, 2020.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[18] Sergio Guadarrama, Ryan Dahl, David Bieber, Mohammad Norouzi, Jonathon Shlens, and Kevin Murphy. Pixcolor: Pixel recursive colorization, 2017.

[19] Raj Kumar Gupta, Alex Yong-Sang Chia, Deepu Rajan, Ee Sin Ng, and Huang Zhiyong. Image colorization using similar images. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 369–378, 2012.

[20] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022.

[21] David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, pages 87–95. SPIE, 2003.

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

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[25] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[26] Yifei Huang, Sheng Qiu, Changbo Wang, and Chenhui Li. Learning representations for high-dynamic-range image color transfer in a self-supervised way. *IEEE Transactions on Multimedia*, 23:176–188, 2020.

[27] Zhitong Huang, Nanxuan Zhao, and Jing Liao. Unicolor: A unified framework for multi-modal colorization with transformer. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022.

[28] 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. In *SIGGRAPH*, pages 110:1–110:11, 2016.

[29] Revital Ironi, Daniel Cohen-Or, and Dani Lischinski. Colorization by example. *Rendering techniques*, 29:201–210, 2005.

[30] Jennifer Jang and Heinrich Jiang. Dbscan++: Towards fast and scalable density clustering, 2019.

[31] 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 *Computer Vision – ECCV 2022*, pages 20–36, Cham, 2022. Springer Nature Switzerland.

[32] Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, and Xuansong Xie. Ddcolor: Towards photorealistic image colorization via dual decoders. *arXiv preprint arXiv:2212.11613*, 2022.

[33] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[34] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[35] 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 *European Conference on Computer Vision (ECCV)*, 2022.

[36] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

[37] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[38] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. *arXiv preprint arXiv:2106.00132*, 2021.

[39] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. *arXiv preprint arXiv:2102.04432*, 2021.

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

[41] Chenyang Lei and Qifeng Chen. Fully automatic video colorization with self-regularization and diversity. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[42] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. In *ACM SIGGRAPH 2004 Papers*, page 689–694, New York, NY, USA, 2004. Association for Computing Machinery.

[43] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity, 2023.

[44] Haoxuan Li, Bin Sheng, Ping Li, Riaz Ali, and CL Philip Chen. Globally and locally semantic colorization via exemplar-based broad-gan. *IEEE Transactions on Image Processing*, 30:8526–8539, 2021.

[45] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[46] Ke Li, Tianhao Zhang, and Jitendra Malik. Diverse image synthesis from semantic layouts via conditional imle. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4220–4229, 2019.

[47] Zekun Li, Zhengyang Geng, Zhao Kang, Wenyu Chen, and Yibo Yang. Eliminating gradient conflict in reference-based line-art colorization, 2022.

[48] Xiaopei Liu, Liang Wan, Yingge Qu, Tien-Tsin Wong, Stephen Lin, Chi-Sing Leung, and Pheng-Ann Heng. Intrinsic colorization. *ACM Trans. Graph.*, 27(5), 2008.

[49] Peng Lu, Jinbei Yu, Xujun Peng, Zhaoran Zhao, and Xiaojie Wang. Gray2colornet: Transfer more colors from reference image. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 3210–3218, New York, NY, USA, 2020. Association for Computing Machinery.

[50] Qing Luan, Fang Wen, Daniel Cohen-Or, Lin Liang, Ying-Qing Xu, and Heung-Yeung Shum. Natural image colorization. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 309–320, 2007.

[51] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023.

[52] Varun Manjunatha, Mohit Iyyer, Jordan Boyd-Graber, and Larry Davis. Learning to color from language, 2018.

[53] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[54] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[55] Yingge Qu, Tien-Tsin Wong, and Pheng-Ann Heng. Manga colorization. *ACM Trans. Graph.*, 25(3):1214–1220, 2006.

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[57] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[58] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

[59] Robin San-Roman, Eliya Nachmani, and Lior Wolf. Noise estimation for generative diffusion models. *arXiv preprint arXiv:2104.02600*, 2021.

[60] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2017.

[61] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[62] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[63] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[64] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[65] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10531–10540, 2019.

[66] Wei Sun and Tianfu Wu. Learning layout and style reconfigurable gans for controllable image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5070–5087, 2021.

[67] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*, 2020.

[68] Yu-Wing Tai, Jiaya Jia, and Chi-Keung Tang. Local color transfer via probabilistic segmentation by expectation-maximization. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 747–754. IEEE, 2005.

[69] Patricia Vitoria, Lara Raad, and Coloma Ballester. Chromagan: Adversarial picture colorization with semantic class distribution, 2020.

[70] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[71] Tomihisa Welsh, Michael Ashikhmin, and Klaus Mueller. Transferring color to greyscale images. 21(3):277–280, 2002.

[72] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. Ct2: Colorization transformer via color tokens. In *ECCV*, 2022.

[73] Shuchen Weng, Hao Wu, Zheng Chang, Jiajun Tang, Si Li, and Boxin Shi. L-code: Language-based colorization using color-object decoupled conditions. In *AAAI*, 2022.

[74] Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. Towards vivid and diverse image colorization with generative color prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[75] Menghan Xia, Wenbo Hu, Tien-Tsin Wong, and Jue Wang. Disentangled image colorization via global anchors. *ACM Transactions on Graphics (TOG)*, 41(6):204:1–204:13, 2022.

[76] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

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

[78] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. Stylization-based architecture for fast deep exemplar colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9363–9372, 2020.

[79] Liron Yatziv and Guillermo Sapiro. Fast image and video colorization using chrominance blending. *IEEE transactions on image processing*, 15(5):1120–1129, 2006.

[80] Wang Yin, Peng Lu, Zhaoran Zhao, and Xujun Peng. Yes," attention is all you need", for exemplar based colorization. In *Proceedings of the 29th ACM international conference on multimedia*, pages 2243–2251, 2021.

[81] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10100–10107. IEEE, 2019.

[82] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[83] Lvmin Zhang, Chengze Li, Tien-Tsin Wong, Yi Ji, and Chunping Liu. Two-stage sketch colorization. *ACM Transactions on Graphics (SIGGRAPH Asia 2018 issue)*, 37(6):261:1–261:14, 2018.

[84] Lvmin Zhang, Chengze Li, Edgar Simo-Serra, Yi Ji, Tien-Tsin Wong, and Chunping Liu. User-guided line art flat filling with split filling mechanism. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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

[86] 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. *ACM Transactions on Graphics (TOG)*, 9(4), 2017.

[87] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[88] Hengyuan Zhao, Wenhao Wu, Yihao Liu, and Dongliang He. Color2embed: Fast exemplar-based image colorization using color embeddings, 2021.

[89] Jiaojiao Zhao, Jungong Han, Ling Shao, and Cees G. M. Snoek. Pixelated semantic colorization, 2019.

[90] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once, 2023.