

Towards Vivid and Diverse Image Colorization with Generative Color Prior

Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, Ying Shan
Applied Research Center (ARC), Tencent PCG

{yanzewu, xintaowang, ianyli, honlanzhang, emmaxunzhao, yingsshan}@tencent.com

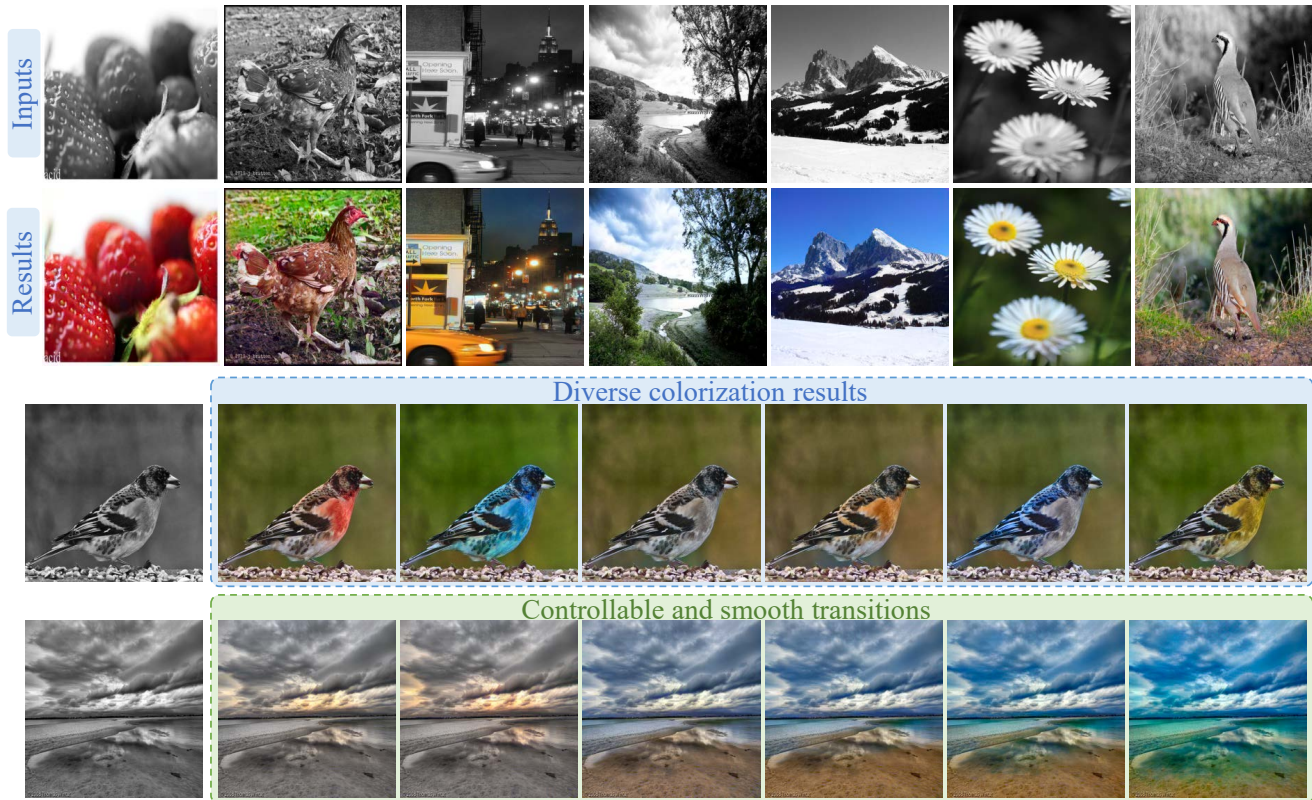


Figure 1: By leveraging the generative color prior and the delicate designs, our method is capable of generating vivid and diverse colorization results without relying on external exemplars. Besides, our method could attain controllable and smooth transitions by walking through the GAN latent space.

Abstract

Colorization has attracted increasing interest in recent years. Classic reference-based methods usually rely on external color images for plausible results. A large image database or online search engine is inevitably required for retrieving such exemplars. Recent deep-learning-based methods could automatically colorize images at a low cost. However, unsatisfactory artifacts and incoherent colors are always accompanied. In this work, we aim at recovering vivid colors by leveraging the rich and diverse color priors encapsulated in a pretrained Generative Adversarial Networks (GAN). Specifically, we first “retrieve” matched features (similar to exemplars) via a GAN encoder and then incorporate these features into the colorization process with

feature modulations. Thanks to the powerful generative color prior and delicate designs, our method could produce vivid colors with a single forward pass. Moreover, it is highly convenient to obtain diverse results by modifying GAN latent codes. Our method also inherits the merit of interpretable controls of GANs and could attain controllable and smooth transitions by walking through GAN latent space. Extensive experiments and user studies demonstrate that our method achieves superior performance than previous works.

1. Introduction

Colorization, the task of restoring colors from black-and-white photos, has wide applications in various fields, such as photography technologies, advertising or film in-

dustry [21, 48]. As colorization requires estimating missing color channels from only one grayscale value, it is inherently an ill-posed problem. Moreover, the plausible solutions of colorization are not unique (e.g., cars in black, blue, or red are all feasible results). Due to the uncertainty and diversity nature of colorization, it remains a challenging task.

Classic reference-based methods (e.g. [17, 35, 53]) require additional example color images as guidance. These methods attempt to match the relevant contents between exemplars and input gray images, and then transfer the color statistics from the reference to the gray one. The quality of generated colors (e.g., realness and vividness) is strongly dependent on the reference images. However, retrieving desirable reference images requires significant user efforts. A recommendation system for this can be a solution, but it is still challenging to design such a retrieval process. A large-scale color image database or online search engine is inevitably required in the system.

Recently, convolutional neural network (CNN) based colorization methods [7, 11, 23] have been proposed to facilitate the colorization task in an automatic fashion. These methods learn to discover the semantics [31, 58] and then directly predict the colorization results. Though remarkable achievements in visual qualities are achieved, unsatisfactory artifacts and incoherent colors are always accompanied.

In this work, we attempt to exploit the merits of both reference-based and CNN-based methods, *i.e.*, achieving realistic and vivid colorization results on par with reference-based methods, while keeping them automatic at a low cost. Inspired by recent success in Generative Adversarial Network (GAN) [3, 27], our core idea is to leverage the most relevant image in the learned GAN distribution as the exemplar image. The rich and diverse color information encapsulated in pretrained GAN models, *i.e.*, generative color prior, allows us to circumvent the explicit example retrieval step and integrate it into the colorization pipeline as a unified framework. Specifically, we first ‘retrieve’ matched features (similar to exemplars) via a GAN encoder and then incorporate these features into the colorization process with feature modulations. In addition, our method is capable of achieving diverse colorization from different samples in the GAN distribution or by modifying GAN latent codes. Thanks to the interpretable controls of GANs, our method could also attain controllable and smooth transitions by walking through GAN latent space.

We summarize the contributions as follows. (1) We develop a unified framework to leverage rich and diverse generative color prior for automatic colorization. (2) Our method allows us to obtain diverse outputs. Controllable and smooth transitions can be achieved by manipulating the code in GAN latent space. (3) Experiments show that our method is capable of generating more vivid and diverse colorization results than previous works.

2. Related Work

User assisted colorization. Early colorization methods are mostly interactive and require users to draw color strokes on the gray image to guide the colorization. The propagation from local hints to all pixels is usually formulated in a constrained optimization manner [33] that tries to assign two pixels with the same color if they are adjacent and similar under similarity measures. Attempts in this direction focus on finding efficient propagation [55] with different similarity metrics by hand-crafting [43] or learning [12], building long-range relations [52], or using edge detection to reduce bleeding [19]. While these methods can generate appealing results with careful interactions, it demands intensive user efforts. The work of [59] alleviates this issue by using sparse color points, and introduces a neural network to colorize from these hints. Some methods also propose to use global hints like color palettes [2, 5] instead of dense color points as constraints.

Reference-based methods try to transfer the color statistics from the reference to the gray image using correspondences between the two based on low-level similarity measures [34], semantic features [6], or super-pixels [8, 15]. Recent works [17, 35, 53] adopt deep neural network to improve the spatial correspondence and colorization results. Though these methods obtain remarkable results when suitable references are available, the procedure of finding references is time-consuming and challenging for automatic retrieval system [8].

Automatic colorization methods are generally learning based [7, 11, 23, 30]. Pre-trained networks for classification are used for better semantic representation [31]. Two branch dual-task structures are also proposed in [22, 48, 60, 61] that jointly learn the pixel embedding and local (semantic maps) or global (class labels) information. The recent work [47] investigates the instance-level features to model the appearance variations of objects. To improve the diversity of generated colors, Zhang *et al.* [58] uses a class-rebalancing scheme, while some methods propose to predict per-pixel color distributions [10, 39, 44] instead of a single color. Yet, their results typically suffer from visual artifacts like unnatural and incoherent colors.

Generative priors of pretrained GANs [3, 26, 27, 28] is previously exploited by GAN inversion [1, 13, 38, 41, 62, 63], which aims to find the closest latent codes given an input image. In colorization, they first ‘invert’ the grayscale image back to a latent code of the pretrained GAN, and then conduct iterative optimization to reconstruct images. Among them, mGANprior [13] attempts to optimize multiple codes to improve the reconstruction quality. In order to reduce the gap between training and testing images, DGP [41] further jointly finetunes the generator and the latent codes. However, these results struggle to faithfully retain the local details, as the low-dimension latent codes

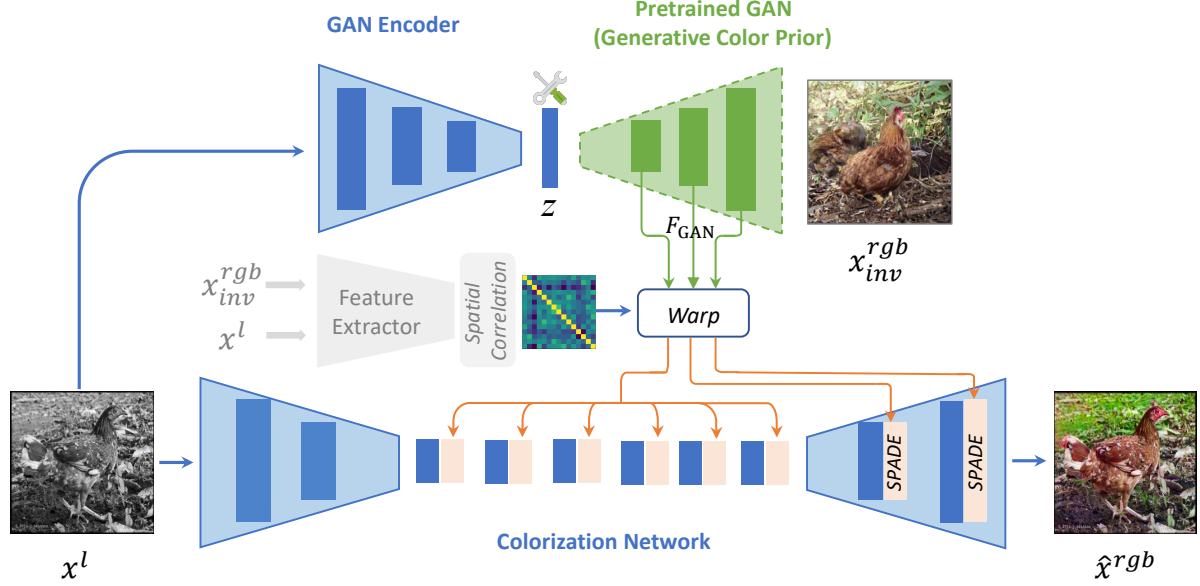


Figure 2: **Overview of our framework.** Given a grayscale image x^l as input, our framework first produces the most relevant features F_{GAN} and an inversion image $x_{\text{inv}}^{\text{rgb}}$ from a pretrained generative network as generative color priors. After that, we calculate a spatial correlation matrix from x^l and $x_{\text{inv}}^{\text{rgb}}$, and warp GAN features F_{GAN} for alignment. The warped features are used to modulate the colorization network by spatially-adaptive denormalization (SPADE) layers. Finally, a vivid colorized image can be generated with the colorization network. In addition, controllable and diverse colorization results, together with smooth transitions could be achieved by adjusting latent code z .

without spatial information are insufficient to guide the colorization. In contrast, our method with spatial modulation enables prior incorporation on multi-resolution spatial features to achieve high texture faithfulness. In addition, our method is feed-forward and does not require expensive iterative optimization for each instance. Recent works on face restoration [51, 54] also utilize generative priors for restoration and colorization.

3. Method

3.1. Overview

With a grayscale image $x^l \in \mathbb{R}^{1 \times H \times W}$ as input, the objective of colorization is to predict the two missing color channels a and b in the original color image $x^{\text{lab}} \in \mathbb{R}^{3 \times H \times W}$, where l and a, b represent the luminance and chrominance in CIELAB color space, respectively. In this work, we aim to leverage rich and diverse color priors encapsulated in a pretrained generative network to guide the colorization. The overview of our framework is shown in Figure 2. It mainly consists of a *pretrained GAN* \mathcal{G} , a *GAN Encoder* \mathcal{E} , and a *Colorization Network* \mathcal{C} .

Given the grayscale image x^l , *GAN Encoder* \mathcal{E} is responsible for mapping x^l into a latent code z , which can serve as an input of \mathcal{G} . The *pretrained GAN* \mathcal{G} then generates the most relevant color image (denoted as inversion image $x_{\text{inv}}^{\text{rgb}}$) to the grayscale input. In addition to using the inversion image directly, the prior features F_{GAN} from intermediate layers are more informative and are leveraged

to guide the colorization. However, these ‘retrieved’ features usually do not spatially align with the grayscale input. To solve this, we first perform an alignment step to warp F_{GAN} based on the correspondence between x^l and $x_{\text{inv}}^{\text{rgb}}$. After that, we use the aligned GAN features to modulate the multi-scale spatial features in *Colorization Network* \mathcal{C} with effective SPADE (spatially-adaptive denormalization) layers [42], producing the final colorized image \hat{x}^{rgb} .

3.2. Generative Color Prior and GAN Encoder

Recently, GAN has achieved tremendous advances in unsupervised image generation. Once trained on an adequate number of images, GAN can map arbitrary random noises into realistic and high-resolution images with impressive details and rich colors, which motivates us to employ pretrained GAN models to assist the colorization task. As illustrated above, we need to find a latent code z so that \mathcal{G} can take it as the input and produce the GAN features F_{GAN} from its intermediate layers which are most relevant to x^l . To achieve this goal, we constrain the generated image from code z to have similar content with x^l but contains rich chrominance information. The process of finding a certain z corresponding to a target image is termed as *GAN inversion*. A typical inversion way is to randomly initialize z and optimize it to reduce the difference between the generated image and the target one. However, such a method requires expensive iterative optimization at inference and thus is not applicable for real-world scenarios. Instead, we choose to train a GAN encoder \mathcal{E} to map arbitrary grayscale images

into latent codes, *i.e.*,

$$\begin{aligned} z &= \mathcal{E}(x^l), \\ x_{inv}^{rgb}, F_{GAN} &= \mathcal{G}(z), \end{aligned} \quad (1)$$

where $\mathcal{G}(z)$ simultaneously outputs the inversion image x_{inv}^{rgb} and the intermediate GAN features F_{GAN} .

Discussion. It is well known that the optimization-based inversion methods could generate more accurate inversion images than encoder-based inversion methods [62, 63]. However, in our colorization scenarios, we do not need such precise and high-quality inversion images. A coarse inversion result already contains sufficient color information and the colorization network (illustrated in the next section) will further refine the corresponding inversion features. In addition, we use the GAN features as extra information to guide the colorization, which could provide complementary information to the inversion image.

3.3. Colorization Network

The colorization network \mathcal{C} consists of two downsampling blocks, six residual blocks and two upsampling blocks, which proves to be an effective architecture used in many image-to-image translation tasks [50]. It takes the grayscale image x^l as input and predicts two missing color channels \hat{x}^{ab} . A color image \hat{x}^{lab} can then be obtained by concatenating the luminance channel x^l and the chrominance channels \hat{x}^{ab} . As the two color spaces RGB and LAB can be converted to each other, \hat{x}^{rgb} can also be obtained.

In order to use the prior color features F_{GAN} to guide the colorization, and to better preserve the color information of F_{GAN} , we use spatially-adaptive denormalization (SPADE) [42] to modulate the colorization network. However, F_{GAN} can not be directly used to modulate network \mathcal{C} until they are spatially aligned with the input. Taking the depicted image in Figure 2 as an example, the inversion image x_{inv}^{rgb} contains all the semantic components that appeared in x^l (*i.e.*, the hen, soil and the weeds), but it is not spatially aligned with the input image.

Before we introduce the align operation, it is worth mentioning that the prior color features F_{GAN} are multi-scale features, and the features from a certain scale are used to modulate the layers of the corresponding scale in \mathcal{C} . For simplicity, we only use a specific scale in F_{GAN} to illustrate the align operation, denoted as $F_{GAN}^s \in \mathbb{R}^{C \times H' \times W'}$.

In the alignment module, we first use two feature extractors with a shared backbone (denoted as $\mathcal{F}_{L \rightarrow S}$ and $\mathcal{F}_{RGB \rightarrow S}$, respectively) to project x^l and x_{inv}^{rgb} to a shared feature space S , obtaining the feature maps $\mathcal{F}_{L \rightarrow S}(x^l)$ and $\mathcal{F}_{RGB \rightarrow S}(x_{inv}^{rgb})$. After that, we use a non-local operation [50] to calculate the correlation matrix $\mathcal{M} \in$

$\mathbb{R}^{H'W' \times H'W'}$ between the two feature maps. $\mathcal{M}(u, v)$ denotes the similarity between the input and the GAN features in position u and v . This operation is widely adopted to calculate dense semantic correspondence between two feature maps [32, 56, 57]. Finally, we use the correlation matrix \mathcal{M} to warp F_{GAN}^s and obtain the aligned GAN features at scale s , which are then used to modulate corresponding layers in \mathcal{C} at scale s .

3.4. Objectives

GAN inversion losses. As we have the ground truth colorful images x^{rgb} during training, we can train GAN encoder \mathcal{E} by minimizing the difference between x_{inv}^{rgb} and x^{rgb} . Instead of directly constraining the pixel values of x_{inv}^{rgb} and x^{rgb} , we choose to minimize the discrepancy between their features extracted by the pre-trained discriminator \mathcal{D}^g of \mathcal{G} , as this will make the GAN inversion easier to optimize (especially for BigGAN) [9, 41]. Formally, the discriminator feature loss \mathcal{L}_{inv_ftr} is defined as:

$$\mathcal{L}_{inv_ftr} = \sum_l \left\| \mathcal{D}_l^g(x_{inv}^{rgb}) - \mathcal{D}_l^g(x^{rgb}) \right\|_1, \quad (2)$$

where \mathcal{D}_l^g represents the feature map extracted at l -th layer from \mathcal{D}^g .

To keep z within a reasonable range, we also add a L_2 norm \mathcal{L}_{inv_reg} , otherwise, the GAN inversion with grayscale input will lead to unstable results:

$$\mathcal{L}_{inv_reg} = \frac{1}{2} \|z\|_2. \quad (3)$$

Adversarial loss. We adopt the adversarial loss to help the colorization images looking more realistic. To stabilize the adversarial training, we employ the loss functions in LS-GAN [36]:

$$\begin{aligned} \mathcal{L}_{adv}^{\mathcal{D}} &= \mathbb{E}[(\mathcal{D}^c(x^{lab}) - 1)^2] + \mathbb{E}[(\mathcal{D}^c(\hat{x}^{lab}))^2], \\ \mathcal{L}_{adv}^{\mathcal{G}} &= \mathbb{E}[(\mathcal{D}^c(\hat{x}^{lab}) - 1)^2], \end{aligned} \quad (4)$$

where \mathcal{D}^c is the discriminator to discriminate the colorization images \hat{x}^{lab} generated from \mathcal{C} and color images x^{lab} from real world. \mathcal{C} and \mathcal{D}^c are trained alternatively with $\mathcal{L}_{adv}^{\mathcal{G}}$ and $\mathcal{L}_{adv}^{\mathcal{D}}$, respectively.

Perceptual loss. To make the colorization image perceptual plausible, we use the perceptual loss introduced in [25]:

$$\mathcal{L}_{perc} = \left\| \phi_l(\hat{x}^{rgb}) - \phi_l(x^{rgb}) \right\|_2, \quad (5)$$

where ϕ_l represents the feature map extracted at l -th layer from a pretrained VGG19 network. Here, we set $l = relu5_2$.

Domain alignment loss. To ensure that the grayscale image and inversion image are mapped to a shared feature

space in correlation calculation, we adopt a domain alignment loss [57]:

$$\mathcal{L}_{dom} = \|\mathcal{F}_{L \rightarrow S}(x^l) - \mathcal{F}_{RGB \rightarrow S}(x^{rgb})\|_1. \quad (6)$$

Contextual Loss. The contextual loss [37] can measure the similarity between two non-aligned images. This loss term was originally designed for image translation tasks with non-aligned data, and then extended to exemplar-based colorization [56] to encourage colors in the output image to be close to those in the reference image. In this paper, we also adopt this loss to encourage the colorization image \hat{x}^{rgb} to be relevant to the inversion image x_{inv}^{rgb} . Formally, the contextual loss \mathcal{L}_{ctx} is defined as:

$$\mathcal{L}_{ctx} = \sum_l \omega_l \left(-\log(\text{CX}(\phi_l(\hat{x}^{rgb}), \phi_l(x_{inv}^{rgb}))) \right), \quad (7)$$

where CX denotes the similarity metric between two features and we refer readers to [37] for more details. We use the layer $l = \text{relu}\{3_2, 4_2, 5_2\}$ from the pretrained VGG19 network with weight $\omega_l = \{2, 4, 8\}$ to calculate \mathcal{L}_{ctx} . We set larger loss weights for deeper features to guide the network to focus more on the semantic similarity instead of pixel-wise color closeness.

Full objective. The full objective to train the GAN encoder \mathcal{E} is formulated as

$$\mathcal{L}_{\mathcal{E}} = \lambda_{inv_ftr} \mathcal{L}_{inv_ftr} + \lambda_{inv_reg} \mathcal{L}_{inv_reg}. \quad (8)$$

The full objective to train the colorization network \mathcal{C} is formulated as

$$\begin{aligned} \mathcal{L}_{\mathcal{C}} = & \lambda_{dom} \mathcal{L}_{dom} + \lambda_{perc} \mathcal{L}_{perc} \\ & + \lambda_{ctx} \mathcal{L}_{ctx} + \lambda_{adv} \mathcal{L}_{adv}^G. \end{aligned} \quad (9)$$

The full objective to train the discriminator \mathcal{D}_C for colorization is formulated as

$$\mathcal{L}_{\mathcal{D}_C} = \lambda_{adv} \mathcal{L}_{adv}^D. \quad (10)$$

In the above loss objectives, λ_{inv_ftr} , λ_{inv_reg} , λ_{dom} , λ_{perc} , λ_{ctx} and λ_{adv} are hyper-parameters that control the relative importance of each loss term.

4. Experiments

4.1. Implementation Details

We conduct our experiments on ImageNet [45], a multi-class dataset, with a pretrained BigGAN [3] as the generative color prior. We train our method with the official training set. All the images are resized to 256×256 .

Our training process consists of two stages. In the first stage, we freeze the pretrained BigGAN generator and train a BigGAN encoder with $\mathcal{L}_{\mathcal{E}}$. We utilize the last three layers of BigGAN discriminator to calculate the feature loss,

where λ_{inv_ftr} is set to 1.0 and λ_{inv_reg} is set to 0.0125. We train the model for totally 10 epochs with an initial learning rate of $1e-5$.

In the second stage, the BigGAN encoder is frozen and the rest of networks are trained end-to-end. We set the λ_{perc} to $1e-3$, λ_{adv} to 1.0, λ_{dom} to 10.0 and λ_{ctx} to 0.1. We train the model for totally 10 epochs with an initial learning rate of $1e-4$. In both two stages, we set the batch size to 8 per GPU and decay the learning rate linearly to 0. All the models are trained with the Adam [29] optimizer.

The BigGAN encoder has a similar structure to the BigGAN discriminator, while the class embedding is injected into the BigGAN encoder following the BigGAN generator. To reduce the learnable parameters of $\mathcal{F}_{L \rightarrow S}$ and $\mathcal{F}_{RGB \rightarrow S}$, their backbones are shared. The discriminator for colorization \mathcal{D}^c is a three-scale PatchGAN network [23, 50] with spectral normalization [40].

4.2. Comparisons with Previous Methods

To evaluate the performance of our method, we compare our results to other state-of-the-art automatic colorization methods including CIC [58], ChromaGAN [48], DeOldify¹ and InstColorization [47]. We re-implement CIC and ChromaGAN in PyTorch based on the original paper.

Quantitative comparison. We test all the methods on the ImageNet validation set of 50,000 images. The quantitative results on five metrics are reported in Table 1. **Fréchet Inception Score (FID)** [18] measures the distribution similarity between the colorization results and the ground truth color images. Our method achieves the lowest FID, indicating that our method could generate colorization results with better image quality and fidelity. **Colorfulness Score** [16] reflects the vividness of generated images. Our method obtains a better colorfulness score (2nd column) and closest colorfulness score to the ground-truth images (3rd column). The highest colorfulness score of CIC is an outlier, also observed in [56]. This is probably because CIC encourages rare colors in the loss function, and the rare colors are misjudged as vivid colors by this metric. We also provide PSNR and SSIM for reference. However, it is well-known that such pixel-wise measurements may not well reflect the actual performance [4, 17, 39, 47, 48, 60], as plausible colorization results probably diverge a lot from the original color image.

Qualitative Comparison. As shown Figure 3, compared to CIC, ChromaGAN and Deoldify, our results tend to obtain more natural and vivid colors. For instance, in contrast to our vivid green, the broccoli color (Column 5) of the other methods looks unnatural and yellowish. In addition, our method generates better results with regards to consistent hues. In column 1, we can observe inconsistent chromaticities in the results of Deoldify and InstColorization, where

¹<https://github.com/jantic/DeOldify>

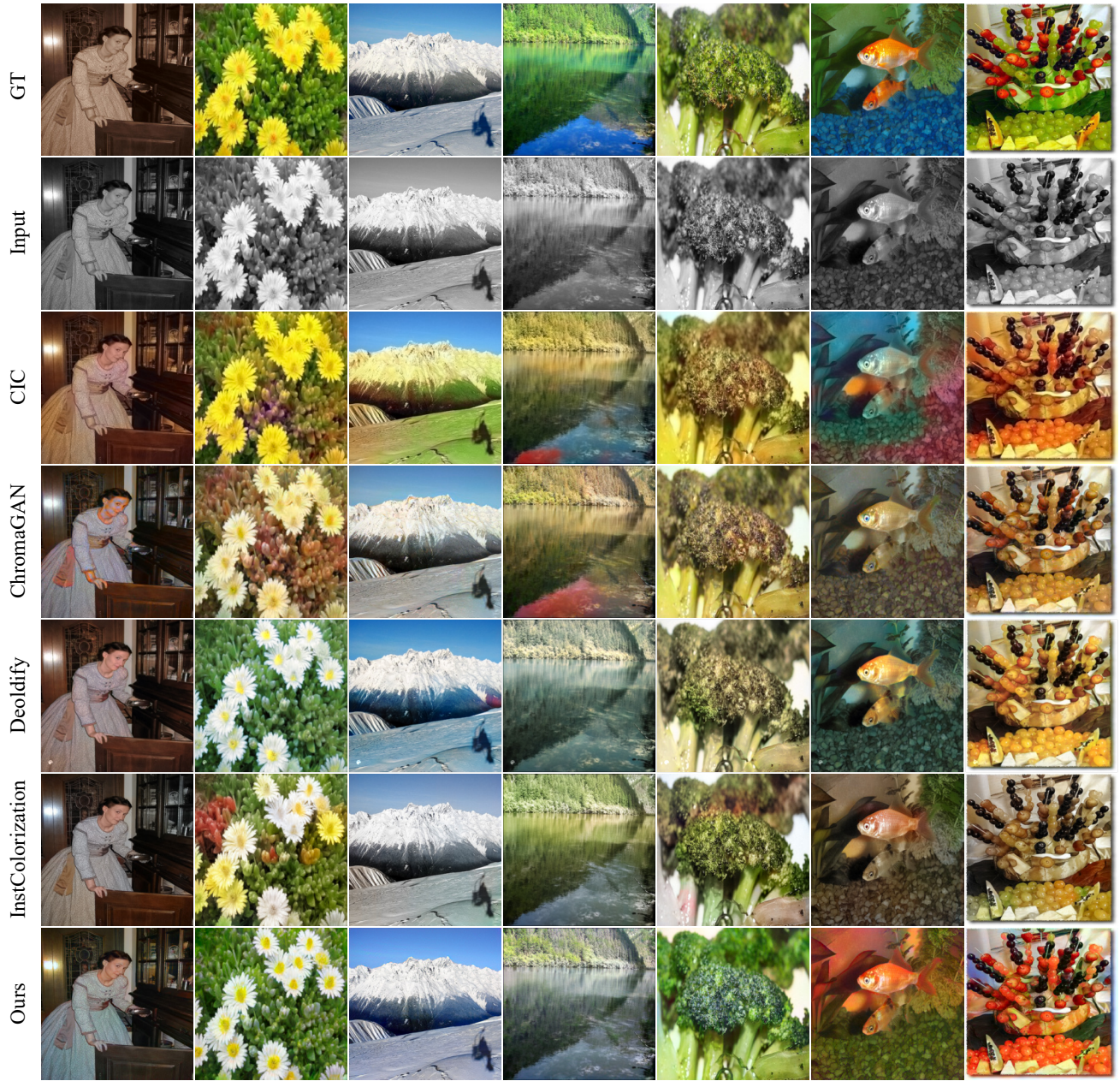


Figure 3: Visual comparisons with previous automatic colorization methods. Our method is able to generate more natural and vivid colors.

Table 1: Quantitative comparison. Δ Colorful denotes the absolute colorfulness score difference between the colorization images and the ground truth color images.

	FID↓	Colorful↑	Δ Colorful↓	PSNR↑	SSIM↑
CIC	19.71	43.92	5.57	20.86	0.86
ChromaGAN	5.16	27.49	10.86	23.12	0.87
DeOldify	3.87	22.83	15.52	22.97	0.91
InstColor	7.36	27.05	11.30	22.91	0.91
Ours	3.62	35.13	3.22	21.81	0.88

the dress is yellow and blue at the same time. Though ChromaGAN has consistent tones, it fails to capture the details by only mapping overall blue tones to the image, leading to the blue artifacts on the lady’s face. Instead, our method successfully maintains the consistent tone and captures the details as shown in Column 1. Furthermore, our method can yield more diverse and lively colors for each small instance as shown in column 7, while InstColorization fails to detect such small objects and inherently degrades to use its global image colorization network, whose results are dim and less vivid.

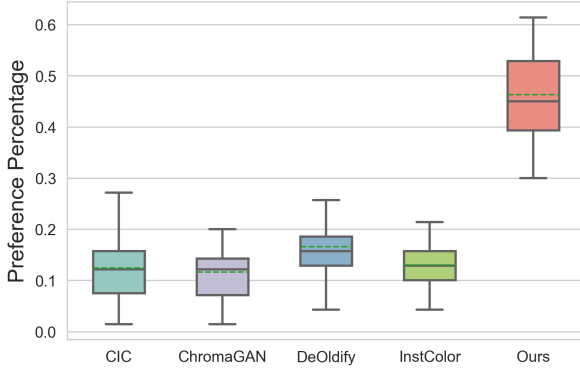


Figure 4: Boxplots of user preferences for different methods. Green dash lines represent the means. Our method got a significantly higher preference rate by users than other colorization methods.

User study. In order to better evaluate the subjective quality (*i.e.*, vividness and diverseness of colors), we conduct a user study to compare our method with the other state-of-art automatic colorization methods, *i.e.*, CIC [58], ChromaGAN [48], DeOldify and InstColorization [47]. We recruited 30 participants with normal or corrected-to-normal vision and without color blindness. We randomly select 70 images from various categories of image contents (*e.g.*, animals, plants, human beings, landscape, food, *etc.*). For each image, the grayscaled image is displayed at the left-most while five colorization results are displayed in a random manner to avoid potential bias. The participants are required to select the best-colored image according to the colorization quality in terms of vividness and diverseness. The boxplots are displayed in Figure 4. Our method obtained a significant preference (46.3%) by users than other colorization methods (CIC 12.4%, ChromaGAN 11.7%, Deoldify 16.6%, InstColor 13.0%), demonstrating distinct advantage on producing natural and vivid results.

4.3. Ablation Studies

Generative color prior. Generative color prior plays an important role in providing exemplar features for vivid and various colors. When we remove the pretrained GANs, our method degrades to a common automatic colorization method without guidance. As shown in Figure 5, the variant experiment without GAN prior generates dim and bleak colors, while our method could produce bright and joyful images, under the guidance of colorful exemplars in GAN distributions. We could also observe a large drop in FID and Colorfulness score without generative color prior (Table 2).

Feature guidance vs. image guidance. Our method adopts multi-resolution intermediate GAN features to guide the colorization. The design could fully facilitate the color and detail information from GAN features. Compared to image guidance, our results are capable of recovering more vivid colors faithful to their underlying classes (*e.g.*, the trees on

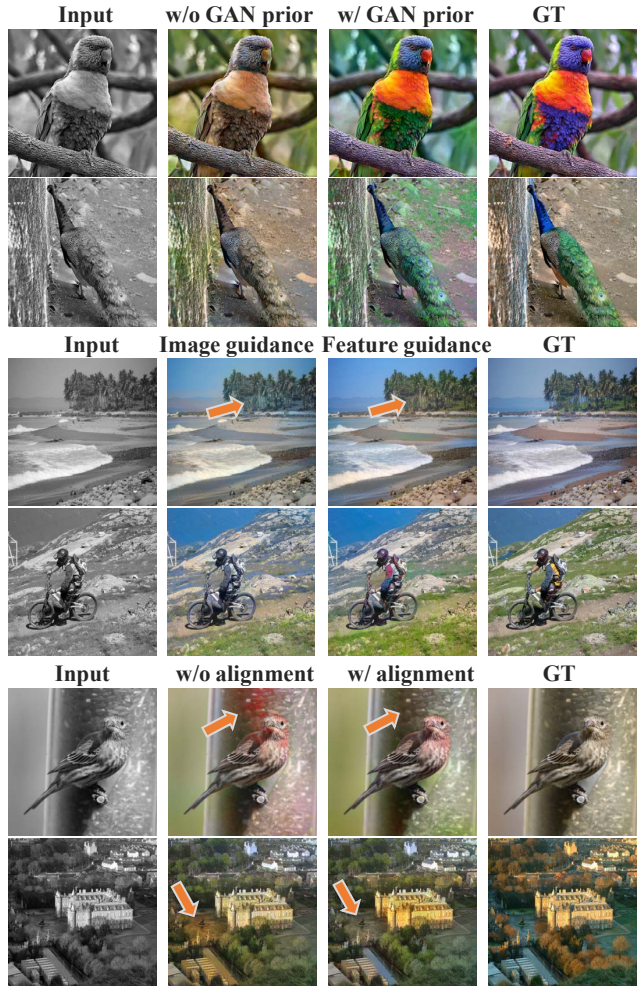


Figure 5: Qualitative comparisons of ablation studies on GAN prior, feature guidance and spatial alignment. GAN prior introduces brighter and more various colors. Multi-resolution feature guidance could produce more coherent and rich colors than image guidance. Spatial alignment mitigates the misalignment between ‘retrieved’ GAN features and input gray images.

the beach, the grass on the mountains), as multi-resolution features contain more semantic and color information (Figure 5 and Table 2).

Spatial alignment. The prior features from pretrained GANs probably misalign with the input images. To address this problem, we employ spatial alignment to align the GAN features, leading to fewer artifacts and coherent results for colorization. As shown in Figure 5, the tree and square colors of our results are not affected by other objects.

4.4. Controllable Diverse Colorization

In order to achieve diverse colorization, many attempts have been made. For automatic colorization methods, diversity is usually achieved by stochastic sampling [10, 14, 44]. However, the results are barely satisfactory and difficult to

Table 2: Quantitative comparisons for ablation studies. Δ Colorful denotes the absolute colorfulness score difference between the colorization images and the ground truth color images.

Variants	FID↓	Colorful↑	Δ Colorful↓
Full Model	3.62	35.13	3.22
w/o Generative Color Prior	8.40	31.21	7.14
Image Guidance	4.01	26.12	12.23
w/o Spatial Alignment	4.59	31.94	6.41



Figure 6: Our method could adjust the latent codes to obtain various inversion results, thus easily achieving diverse colorization results for the parrot.

control (Detailed comparisons are provided in the supplementary material). For exemplar-based colorization methods [17, 53], it is time-consuming and challenging to find large amounts of reference images with various styles. Different from previous methods, our method could achieve diverse and controllable colorization in a much easier and novel manner.

On the one hand, we can adjust the latent codes with disturbance to obtain various exemplars and their corresponding features. As shown in Figure 6, various inversion results are attained by adjusting the latent codes, and correspondingly, diverse colorization results of parrots are obtained. On the other hand, our method inherits the merits of interpretable controls of GANs [20, 24, 46, 49] and could attain controllable and smooth transitions by walking through GAN latent space. Specifically, we employ an unsupervised method [49] to find color-relevant directions, such as lighting, saturation, *etc.* As depicted in Figure 7, smooth transitions on background colors, object colors, saturation could be achieved.

4.5. Limitations

Though our method could produce appealing results in most cases, it still has limitations. When the input image is not in the GAN distribution or GAN inversion fails, our method degrades to common automatic colorization methods and may result in unnatural and incoherent colors. As shown in Figure 8, the human on the beach and the cactus are missing from the GAN inversion results. Thus, our method cannot find corresponding color guidance, resulting in unnatural colorization for these objects. This could be mitigated by improving the GAN inversion and common automatic colorization approaches in future works.

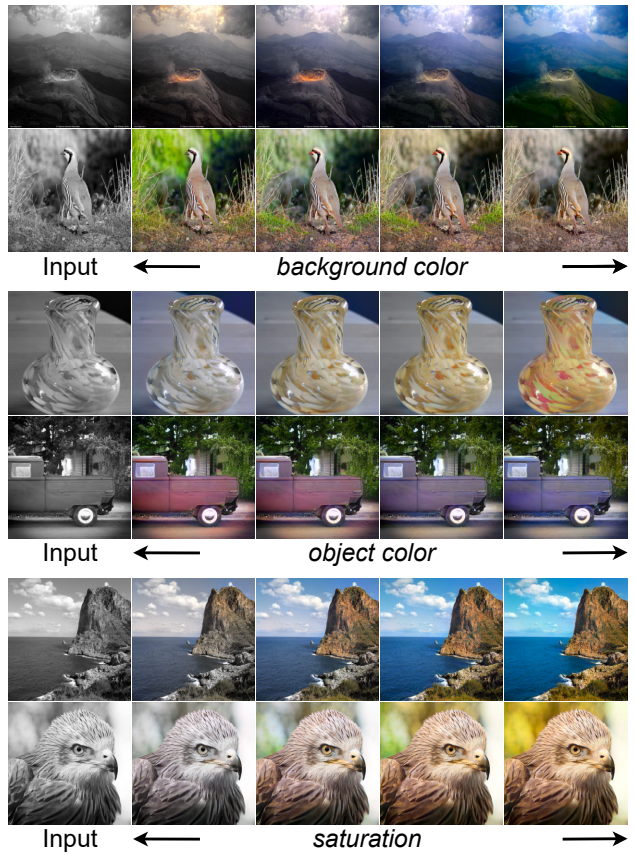


Figure 7: With the interpretable controls of GANs, our method could attain controllable and smooth transitions by walking through GAN latent space.



Figure 8: Limitations of our model. The human in the beach and the cactus are missing from the GAN inversion, resulting in unnatural colors.

5. Conclusion

In this work, we have developed a framework to produce vivid and diverse colorization results by leveraging generative color priors encapsulated in pretrained GANs. Specifically, we ‘retrieve’ multi-resolution GAN features conditioned on the input grayscale image with a GAN encoder, and incorporate such features into the colorization process with feature modulations. One could easily realize diverse colorization by simply adjusting the latent codes or conditions in GANs. Moreover, our method could also attain controllable and smooth transitions by walking through the GAN latent space.

References

- [1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019.
- [2] Hyojin Bahng, Seungjoo Yoo, Wonwoong Cho, David Keetae Park, Ziming Wu, Xiaojuan Ma, and Jaegul Choo. Coloring with words: Guiding image colorization through text-based palette generation. In *ECCV*, 2018.
- [3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- [4] Yun Cao, Zhiming Zhou, Weinan Zhang, and Yong Yu. Unsupervised diverse colorization via generative adversarial networks. In *Joint European conference on machine learning and knowledge discovery in databases*, 2017.
- [5] Huiwen Chang, Ohad Fried, Yiming Liu, Stephen DiVerdi, and Adam Finkelstein. Palette-based photo recoloring. *ACM TOG*, 34(4):139–1, 2015.
- [6] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf. Automatic image colorization via multimodal predictions. In *ECCV*, 2008.
- [7] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *ICCV*, 2015.
- [8] 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. In *SIGGRAPH*, 2011.
- [9] Giannis Daras, Augustus Odena, Han Zhang, and Alexandros G Dimakis. Your local gan: Designing two dimensional local attention mechanisms for generative models. In *CVPR*, 2020.
- [10] Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, and David Forsyth. Learning diverse image colorization. In *CVPR*, 2017.
- [11] Aditya Deshpande, Jason Rock, and David Forsyth. Learning large-scale automatic image colorization. In *ICCV*, 2015.
- [12] Yuki Endo, Satoshi Iizuka, Yoshihiro Kanamori, and Jun Mitani. Deepprop: Extracting deep features from a single image for edit propagation. In *Computer Graphics Forum*, volume 35, pages 189–201, 2016.
- [13] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, 2020.
- [14] Sergio Guadarrama, Ryan Dahl, David Bieber, Jonathon Shlens, Mohammad Norouzi, and Kevin Murphy. PixColor: Pixel recursive colorization. In *BMVC*, 2017.
- [15] Raj Kumar Gupta, Alex Yong-Sang Chia, Deepu Rajan, Ee Sin Ng, and Huang Zhiyong. Image colorization using similar images. In *ACM MM*, 2012.
- [16] David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pages 87–95. International Society for Optics and Photonics, 2003.
- [17] Mingming He, Dongdong Chen, Jing Liao, Pedro V. Sander, and Lu Yuan. Deep exemplar-based colorization. *ACM TOG*, 37(4):1–16, 2018.
- [18] 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. In *NeurIPS*, 2017.
- [19] Yi-Chin Huang, Yi-Shin Tung, Jun-Cheng Chen, Sung-Wen Wang, and Ja-Ling Wu. An adaptive edge detection based colorization algorithm and its applications. In *ACM MM*, 2005.
- [20] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *NeurIPS*, 2020.
- [21] Satoshi Iizuka and Edgar Simo-Serra. Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM TOG*, 38(6):1–13, 2019.
- [22] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color! *ACM TOG*, 35(4):1–11, 2016.
- [23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.
- [24] Ali Jahanian, Lucy Chai, and Phillip Isola. On the “steerability” of generative adversarial networks. In *ICLR*, 2020.
- [25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.
- [26] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.
- [27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2018.
- [28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020.
- [29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- [30] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. In *ICLR*, 2021.
- [31] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *ECCV*, 2016.
- [32] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In *CVPR*, 2020.
- [33] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. *ACM TOG*, 23(3):689–694, 2004.
- [34] Xiaopei Liu, Liang Wan, Yingge Qu, Tien-Tsin Wong, Stephen Lin, Chi-Sing Leung, and Pheng-Ann Heng. Intrinsic colorization. In *SIGGRAPH*, 2008.
- [35] Peng Lu, Jinbei Yu, Xujun Peng, Zhaoran Zhao, and Xiaojie Wang. Gray2colonet: Transfer more colors from reference image. In *ACM MM*, 2020.
- [36] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017.
- [37] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *ECCV*, 2018.

- [38] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020.
- [39] Safa Messaoud, David Forsyth, and Alexander G. Schwing. Structural consistency and controllability for diverse colorization. In *ECCV*, 2018.
- [40] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.
- [41] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. In *ECCV*, 2020.
- [42] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019.
- [43] Yingge Qu, Tien-Tsin Wong, and Pheng-Ann Heng. Manga colorization. *ACM TOG*, 25(3):1214–1220, 2006.
- [44] Amelie Royer, Alexander Kolesnikov, and Christoph Lampert. Probabilistic image colorization. In *BMVC*, 2017.
- [45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.
- [46] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021.
- [47] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *CVPR*, 2020.
- [48] Patricia Vitoria, Lara Raad, and Coloma Ballester. Chromagan: Adversarial picture colorization with semantic class distribution. In *WACV*, 2020.
- [49] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. 2020.
- [50] 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 *CVPR*, 2018.
- [51] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021.
- [52] Kun Xu, Yong Li, Tao Ju, Shi-Min Hu, and Tian-Qiang Liu. Efficient affinity-based edit propagation using kd tree. *ACM TOG*, 28(5):1–6, 2009.
- [53] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. Stylization-based architecture for fast deep exemplar colorization. In *CVPR*, 2020.
- [54] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, 2021.
- [55] L. Yatziv and G. Sapiro. Fast image and video colorization using chrominance blending. *IEEE TIP*, 15(5):1120–1129, 2006.
- [56] Bo Zhang, Mingming He, Jing Liao, Pedro V. Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *CVPR*, 2019.
- [57] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *CVPR*, 2020.
- [58] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *ECCV*, 2016.
- [59] 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 TOG*, 36(4):1–11, 2017.
- [60] Jiaojiao Zhao, Jungong Han, Ling Shao, and Cees G. M. Snoek. Pixelated semantic colorization. *IJCV*, 128(4):818–834, 2019.
- [61] Jiaojiao Zhao, Li Liu, Cees GM Snoek, Jungong Han, and Ling Shao. Pixel-level semantics guided image colorization. In *BMVC*, 2018.
- [62] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *ECCV*, 2020.
- [63] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016.