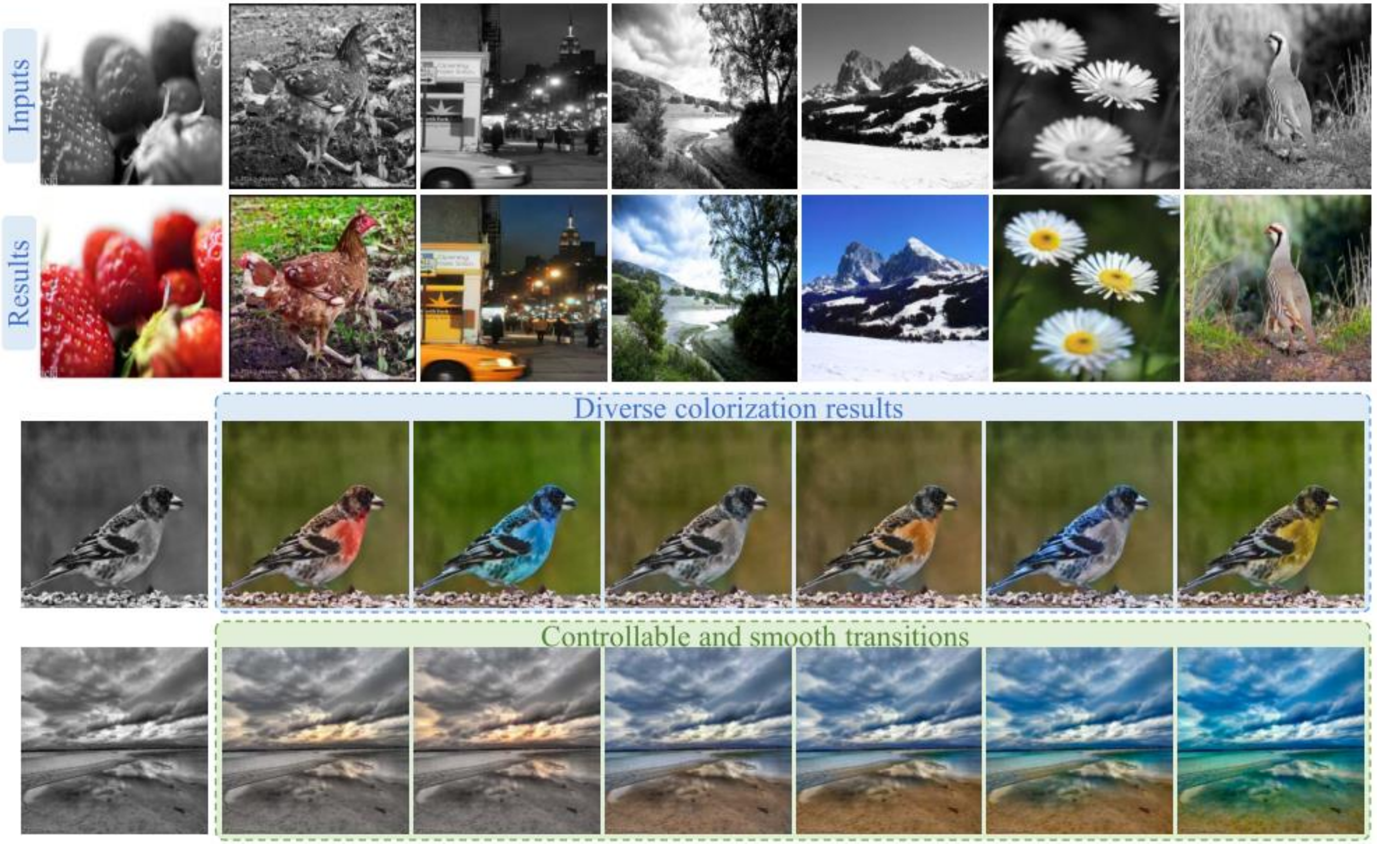
Towards Vivid and Diverse Image Colorization with Generative Color Prior

ANZE WU, XINTAO WANG, YU LI,

HONGLUN ZHANG, XUN ZHAO, YING SHAN

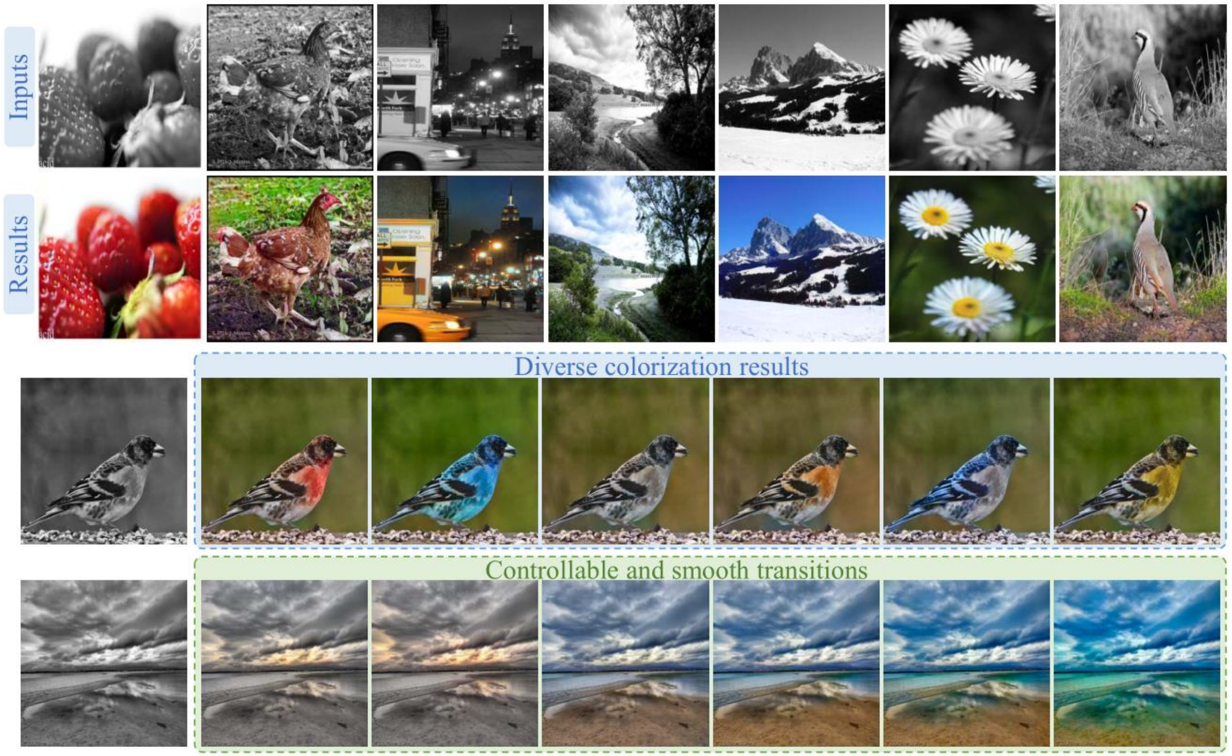
Towards Vivid and Diverse Image Colorization with Generative Color Prior

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



Introduction

Colorization: the task of restoring colors from black-and-white photos



Introduction

Reference-based

- Additional example color images as guidance
- A large-scale color image database or online search engine is inevitably required in the system

CNN based

- Automatic
- Learn to discover the semantics, and then directly predict the colorization results
- Unsatisfactory artifacts and incoherent colors

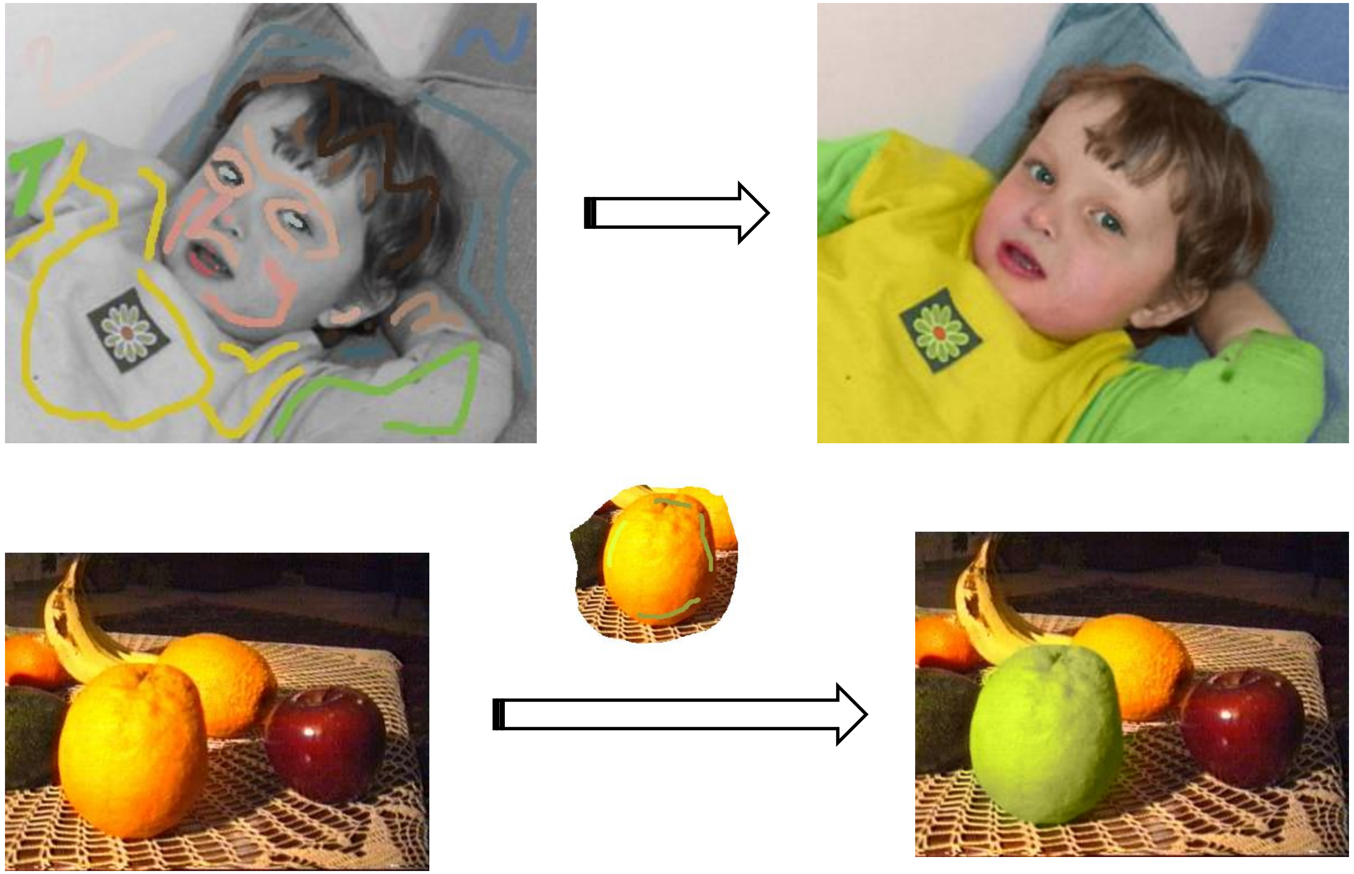
Introduction

- Reference-based + CNN-based method
  - This is a unified framework to leverage rich and diverse generative color prior for automatic colorization.
- Retrieve features via a GAN encoder and then incorporate these features into the colorization process.
- Achieving diverse colorization from different samples in the GAN distribution or by modifying GAN latent codes.

Related Work

User assisted colorization

- Require users to draw color strokes on the gray image to guide the colorization
- Assign two pixels with the same color if they are adjacent and similar under similarity measures.

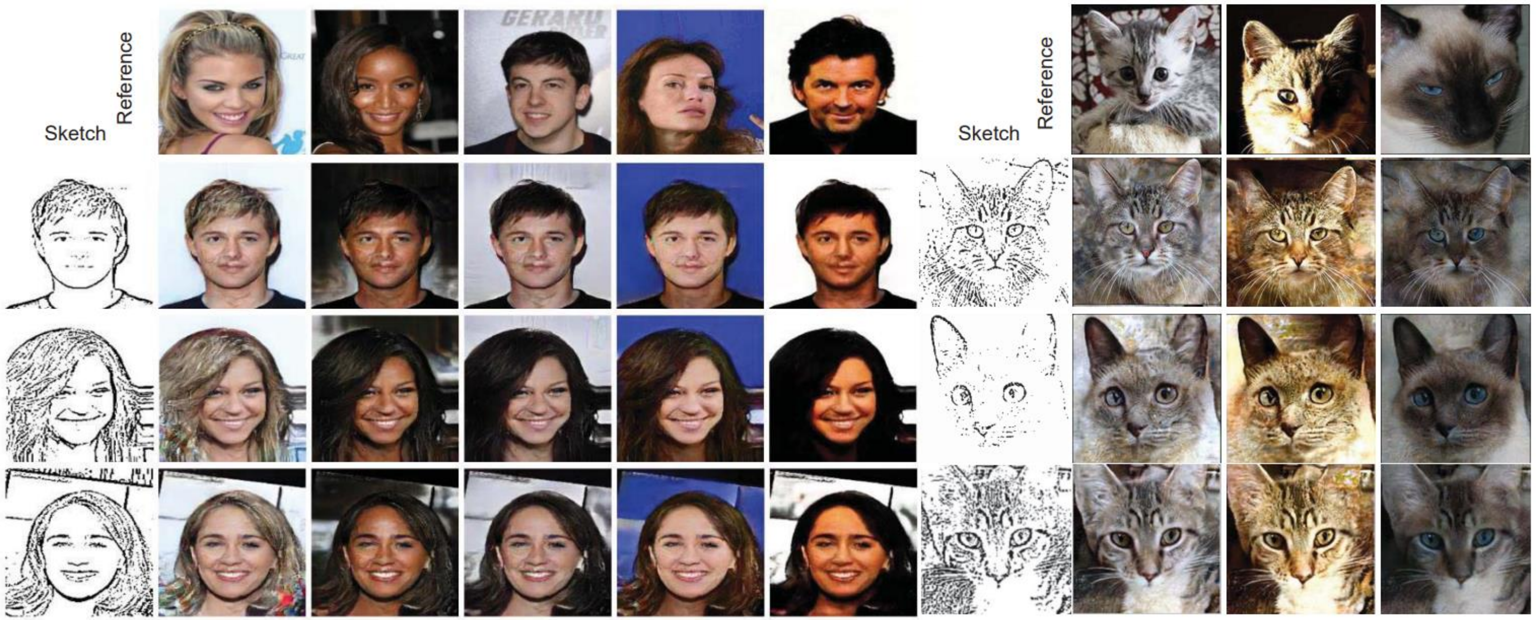


https://www.cs.huji.ac.il/w~yweiss/Colorization/

Related Work

Reference-based methods

- Transfer the color statistics from the reference to the gray image using correspondences between the two based on
  - low-level similarity measures
  - semantic features
  - super-pixels
- The procedure of finding references is time-consuming and challenging for automatic retrieval system



Related Work

Automatic colorization

- Pre-trained networks for classification are used for better semantic representation.[1]
- Two branch dual-task structures are also proposed [2] in that jointly learn the pixel embedding and local (semantic maps) or global (class labels) information
- The recent work [3] investigates the instance-level features to model the appearance variations of objects

1. Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In ECCV, 2016
2. Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color! ACM TOG, 35(4):1–11, 2016.
3. Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In CVPR, 2020.

Related Work Generative priors

- Generative priors of pretrained GANs is exploited by GAN inversion 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
- However, these results struggle to faithfully retain the local details, as the low-dimension latent codes without spatial information are insufficient to guide the colorization.

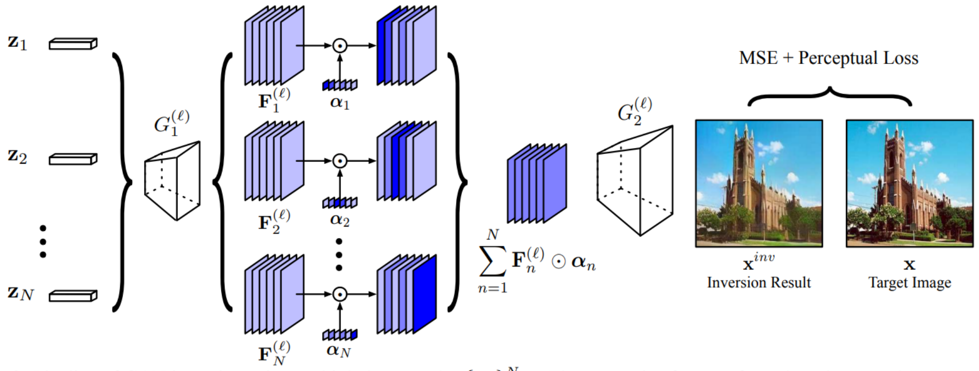


Figure 2: Pipeline of GAN inversion using multiple latent codes $\{\mathbf{z}_n\}_{n=1}^N$. The generative features from these latent codes are composed at some intermediate layer (*i.e.*, the ℓ-th layer) of the generator, weighted by the adaptive channel importance scores $\{\boldsymbol{\alpha}_n\}_{n=1}^N$. All latent codes and the corresponding channel importance scores are jointly optimized to recover a target image.

Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In CVPR, 2020

Method

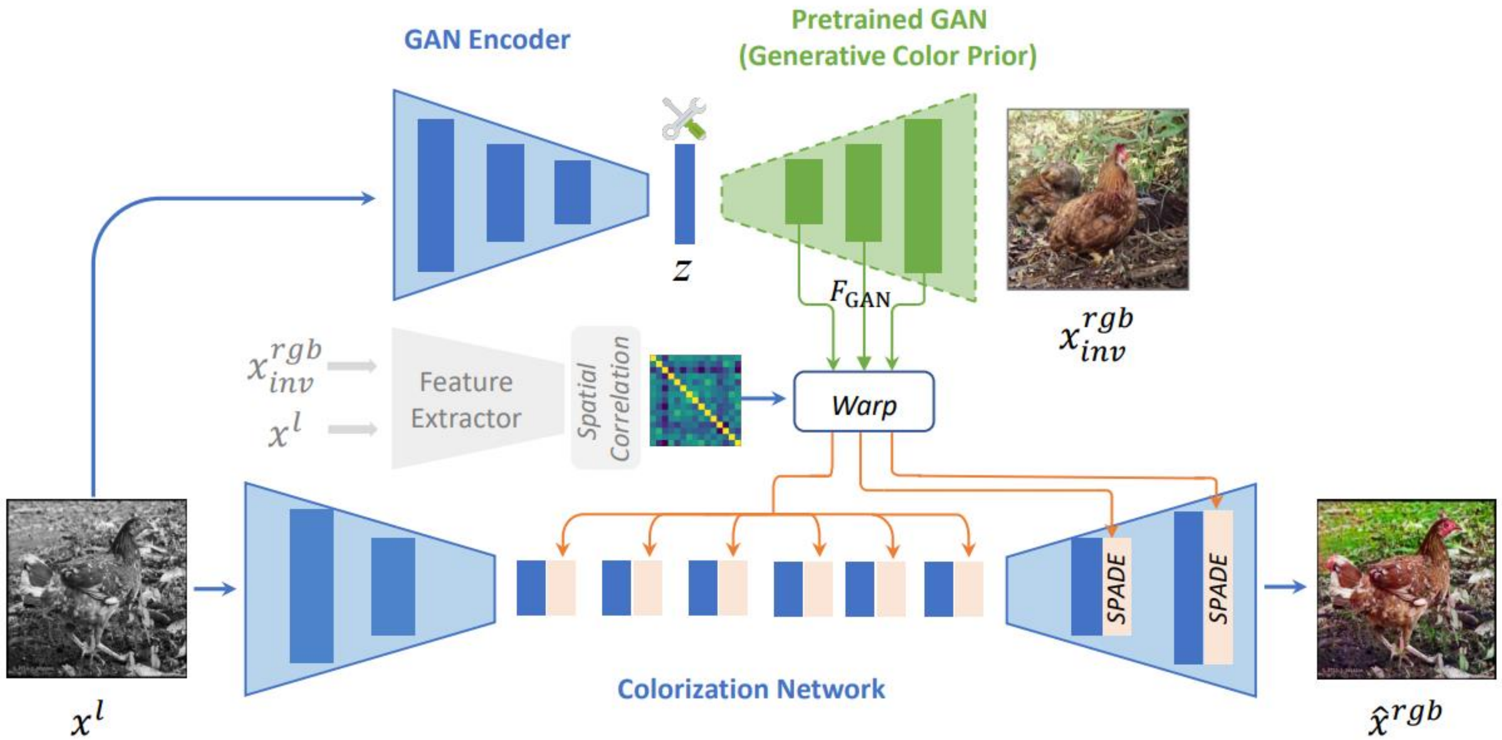
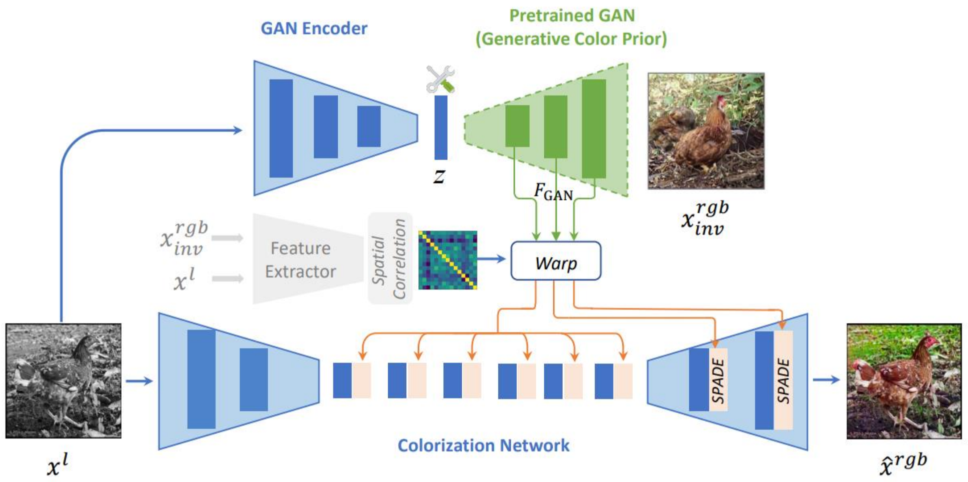


Figure 2: **Overview of our framework**. Given a grayscale image x^l as input, our framework first produces the most relevant features F_{GAN} and an inversion image x_{inv}^{rgb} from a pretrained generative network as generative color priors. After that, we calculate a spatial correlation matrix from x^l and x_{inv}^{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.

Colorization Network



- 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) to modulate the colorization network.
  - 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.
  - we first use two feature extractors with a shared backbone (denoted as $F_{L\to S}$ and $F_{RGB\to S}$, respectively) to project x_l and x_{inv}^{rgb} to a shared feature space S, obtaining the feature maps $F_{L\to S}(x_l)$ and $F_{RGB\to S}(x_{inv}^{rgb})$. After that, we use a non-local operation to calculate the correlation matrix M
  - Finally, we use the correlation matrix M to warp F_{GAN}^{s} and obtain the aligned GAN features at scale s, which are then used to modulate corresponding layers in C at scale s.

Objectives

GAN inversion losses

- We choose to minimize the discrepancy between x_{inv}^{rgb} and x^{rgb} features extracted by the pre-trained discriminator D^{g} of Pretrained GAN
- $L_{inv-ftr} = \sum_{l} \left|\left|D_{l}^{g}\left(x_{inv}^{rgb}\right) - D_{l}^{g}\left(x^{rgb}\right)\right|\right|_{1}$
- D_{l}^{g} represents the feature map extracted at l-th layer from D^{g}
- $L_{inv-reg} = \frac{1}{2}\left|\left|z\right|\right|_{2}$

Objectives

Adversarial loss

- $\mathrm{L}_{adv}^{D} = \mathbb{E}\left[\left(D^c\left(x^{lab}\right) - 1\right)^2\right] + \mathbb{E}\left[\left(D^c\left(\hat{x}^{lab}\right)\right)^2\right]$
- $L_{adv}^{G} = \mathbb{E}\left[\left(D^c\left(\hat{x}^{lab}\right) - 1\right)^2\right]$
- where D^c is the discriminator to discriminate the colorization images $\hat{x}^{lab}$ generated from C and color images x^{lab} from real world. C and D^c are trained alternatively with L_{adv}^{G} and L_{adv}^{D}, respectively

Objectives

Perceptual loss

- To make the colorization image perceptual plausible, we use the perceptual loss
- $L_{perc} = \left\| \phi_l(\hat{x}^{lab}) - \phi_l(x^{lab}) \right\|_2$
- where ϕ_l represents the feature map extracted at l-th layer from a pretrained VGG19 network. Here ,we set $l = relu5\_2$.

Objectives

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
- $\mathrm{L}_{\mathrm{dom}} = \left|\left|F_{L\to S}(x_l) - F_{RGB\to S}\left(x_{inv}^{rgb}\right)\right|\right|_1$

Objectives

Contextual Loss

- We use contextual loss to encourage the colorization image $\hat{x}^{rgb}$ to be relevant to the inversion image x_{inv}^{rgb}.
- $L_{ctx} = \sum_l \omega_l (-\log(CX(\phi_l(\hat{x}^{rgb}), \phi_l(x_{inv}^{rgb}))))$
- where CX denotes the similarity metric between two features.
- We use the layer $l = relu\{3\ 2, 4\ 2, 5\_2\}$ from the pretrained VGG19 network with weight $\omega_l = \{2, 4, 8\}$ to calculate L_{ctx}.

Objectives

Full objective

- The full objective to train the GAN encoder is formulated as
  - $L_{\varepsilon} = \lambda_{inv-ftr}\, L_{inv-ftr} + \lambda_{inv-reg}\, L_{inv-reg}$
- The full objective to train the colorization network is formulated as
  - $L_C = \lambda_{\mathrm{dom}}\, L_{\mathrm{dom}} + \lambda_{prec}\, L_{prec} + \lambda_{ctx}\, L_{ctx} + \lambda_{adv}\, L_{adv}^{G}$
- The full objective to train the discriminator DC for colorization is formulated as
  - $L_{D_C} = \lambda_{adv}\, L_{adv}^{G}$
- $\lambda_{inv-ftr}$, $\lambda_{inv-reg}$, λ_{dom}, λ_{prec}, λ_{ctx} and λ_{adv} are hyper-parameters

Experiments

- We conduct our experiments on ImageNet ,a multiclass dataset, with a pretrained BigGAN as the generative color prior. We train our method with the official training set. All the images are resized to 256×256.
  - Freeze the pretrained BigGAN generator and train a BigGAN encoder with L_{ε}
  - The BigGAN encoder is frozen, and the rest of networks are trained end-to-end.

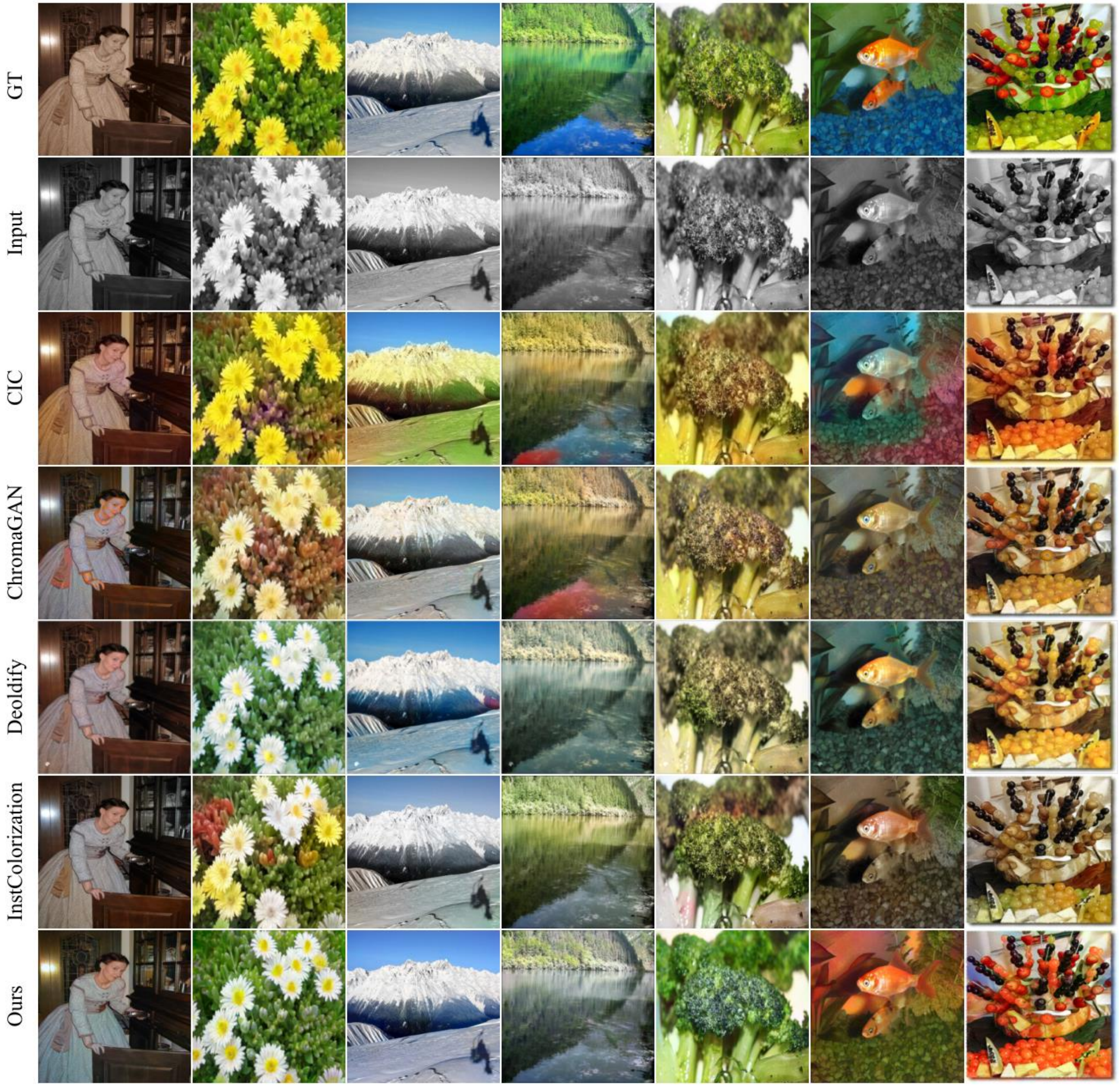
Experiments

Comparisons with Previous Methods

- **Frechet Inception Score (FID) :** measures the distribution similarity between the colorization results and the ground truth color images.
- **Colorfulness Score:** reflects the vividness of generated images.
- **PSNR:** peak signal-to-noise ratio is an expression for the ratio between the maximum possible value of a signal and the power of distorting noise that affects the quality of its representation.
- **SSIM:** The structural similarity index measure is a method for predicting the perceived quality of digital television and cinematic pictures.

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

	FID$\downarrow$	Colorful$\uparrow$	ΔColorful$\downarrow$	PSNR$\uparrow$	SSIM$\uparrow$
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


GT
Input
CIC
ChromaGAN
Deoldify
InstColorization
Ours

Experiments

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

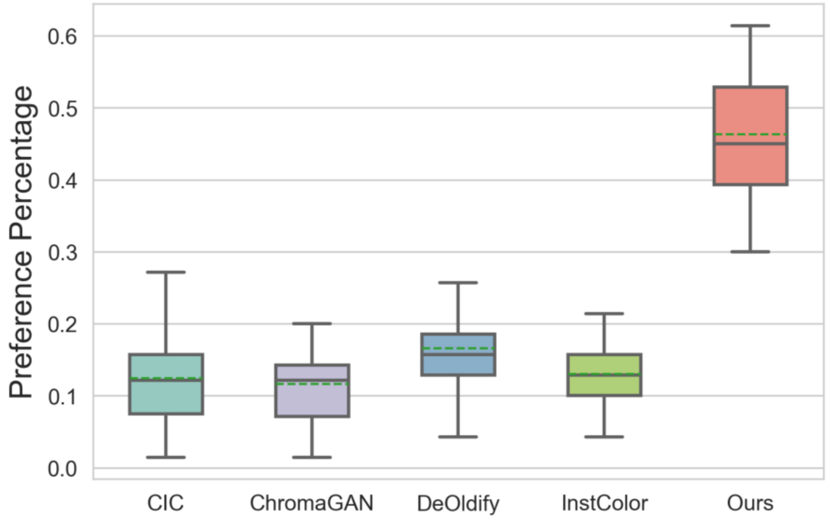


Figure 4: Boxplots of user preferences for different methods. Green dash lines represent the means. Our method got a significantly higher preference rate by users than other colorization methods.

Experiments

Generative color prior

Feature guidance vs. image guidance.

Spatial alignment

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

Variants	FID↓	Colorful↑	ΔColorful↓
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

Experiments

Controllable Diverse Colorization

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

Experiments

Controllable Diverse Colorization

We employ an unsupervised method to find color-relevant directions, such as lighting, saturation, *etc*.

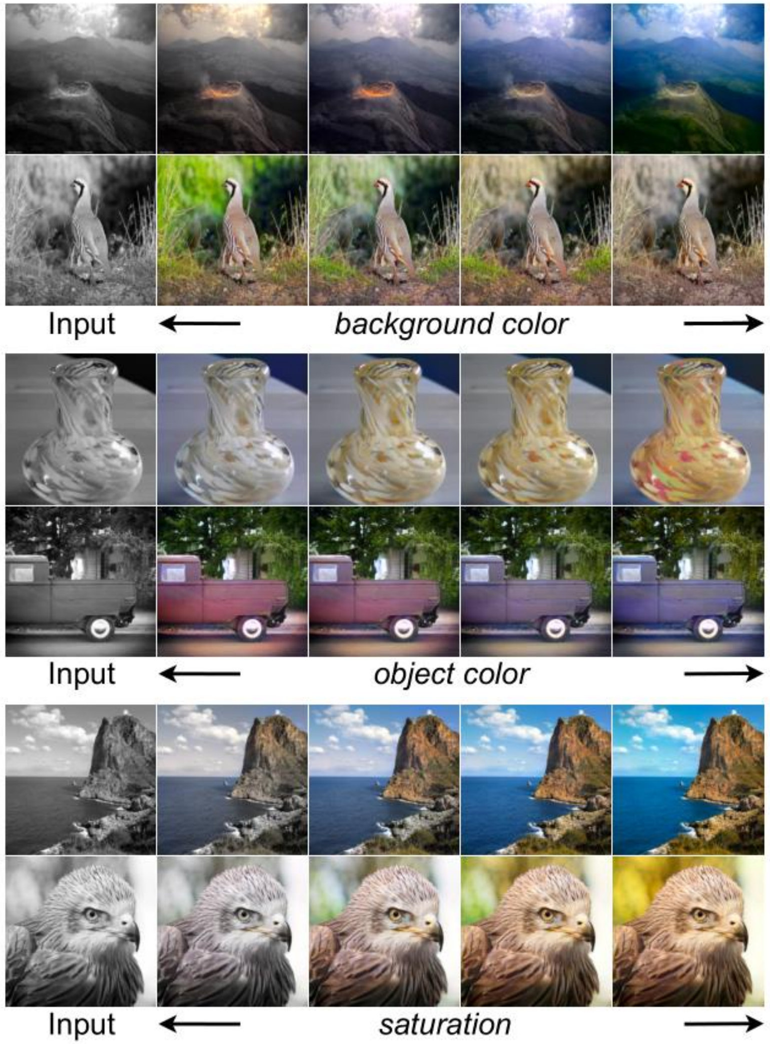


Figure 7: With the interpretable controls of GANs, our method could attain controllable and smooth transitions by

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.

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