UniCoRN: A Unified Conditional Image Repainting Network

Jimeng Sun¹ Shuchen Weng² Zheng Chang¹ Si Li¹ Boxin Shi²,³,⁴
¹School of Artificial Intelligence, Beijing University of Posts and Telecommunications
²National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
³Institute for Artificial Intelligence, Peking University
⁴Beijing Academy of Artificial Intelligence
{sjs, zhengchang98, lisi}@bupt.edu.cn, {shuchenweng, shiboxin}@pku.edu.cn

Abstract

Conditional image repainting (CIR) is an advanced image editing task, which requires the model to generate visual content in user-specified regions conditioned on multiple cross-modality constraints, and composite the visual content with the provided background seamlessly. Existing methods based on two-phase architecture design assume dependency between phases and cause color-image incongruity. To solve these problems, we propose a novel Unified Conditional image Repainting Network (UniCoRN). We break the two-phase assumption in the CIR task by constructing the interaction and dependency relationship between background and other conditions. We further introduce the hierarchical structure into cross-modality similarity model to capture feature patterns at different levels and bridge the gap between visual content and color condition. A new LANDSCAPE-CIR dataset is collected and annotated to expand the application scenarios of the CIR task. Experiments show that UniCoRN achieves higher synthetic quality, better condition consistency, and more realistic compositing effect.

1. Introduction

Advanced image editing is desired in various applications such as colorizing old photos [8, 19, 42, 43], repairing damaged regions [23, 35, 37, 38], blending multiple images [25, 34, 41], and so on. With rapid progress in improving generative networks, skill barriers of using image editing tools have been lowered. For example, users can transform any photo into the style of a “famous painter” simply by providing one of his own works [14].

To “free” the users from requiring professional skills while maintaining the “freedom” to realize their ideas for editing an image, conditional image repainting (CIR) [32, 33] has been proposed. For the CIR task, “repainting” means some regions of an existing image are repainted with new visual content, and “conditional” means such new visual content is generated from cross-modality input conditions which consist of texture (random noise), color (attribute or language), geometry (segmentation mask), and background (RGB image). An example of the CIR task is shown in the top row of Fig. 1.

Existing CIR methods [32, 33] are implemented in a two-
phase manner, as shown in the bottom row of Fig. 1. In the first generation phase, the model generates the visual content under the guidance of input conditions (Fig. 1(a)). In the second compositing phase, it discards the meaningless background regions generated before (the purple area in Fig. 1(b)), and replaces them with the given background to be the input of the compositing model (Fig. 1(c)). After that, the compositing model adjusts the color tone of repainted regions to make the whole image harmonious, and finally synthesizes the repainted image.

Despite that existing two-phase methods could produce reasonable results, there are still several problems. (i) Dependency of two phases: By explicitly dividing the task into two phases, the compositing model can only adjust the color tone of repainted regions after the first generation phase, which limits its “play space” and leads to color tone gaps between regions. Beyond that, along with the discard of meaningless background (the purple area in Fig. 1(b)), the gradient of this area is truncated, causing the gradient backpropagation unstable and further bringing obvious artifacts. (ii) Incongruity between image and color: The cross-modality similarity model (CMSM) [36] has been introduced into two-phase CIR methods to bridge the gap between synthetic images and color conditions. However, it remains to be addressed that the global encoders in CMSM are difficult to provide adequate information, leading to the incongruity between repainted images and input color conditions (MISC synthesizes purple skirt in Fig. 1, while the input color condition is yellow).

In this paper, we propose a Unified Conditional image Repainting Network, denoted as UniCoRN, to solve the above issues in the two-phase CIR methods. Specifically, we redesign the condition fusion and injection modules for the CIR task. By constructing the interaction and dependency relationship between background and other conditions that describe the visual content, we relax the dependency between the generation and compositing phase. Besides, we propose a hierarchical cross-modality similarity model (HCMSM) to extract features at different semantic levels – low-level features are local and coarse-grained while high-level ones are global and fine-grained – to better constrain color consistency in a continuous feature space.

The contributions of this work are two-fold:

- We break the two-phase dependency assumption in the CIR task with a newly designed unified framework, which facilitates conditional image repainting with higher synthetic quality, better condition consistency, and more realistic compositing effect (top right in Fig. 1).
- We collect a high-resolution LANDSCAPE-CIR dataset, including 28K training images, 3K test images, and other necessary inputs obtained by automatic annotations, to expand the application scenarios of the CIR task.

2. Related Works

Conditional generative adversarial networks. Conditional generative adversarial networks (cGANs) are a type of GANs [10], which take special conditions as inputs to constrain the generated results, e.g., using single tags as the class conditions to specify the categories of the generated images [2, 20, 21, 39], using language descriptions to guide image generation [26, 27, 36, 40], using reference images, sketches, or scene graphs for higher control flexibility [13, 17, 18], and converting image-like data to photo-realistic images according to user-given rules [11, 15, 31].

Condition injection. The adaptive instance normalization (AdaIN) [14] is widely used in vector injection and famous for image style transfer. SPatially-Adaptive (DE)normalization (SPADE) [22] is mainly used in image-like data such as segmentation mask, which can better preserve semantic information in uniform or flat regions. The semantic region-adaptive normalization (SEAN) [46] is a simple yet effective module, which constructs a style map under the guidance of segmentation mask and learns element-wise normalization values to control the style of each semantic region individually. Semantic-style block [24] derives attribute or language into 64D representation and concatenates it with class embedding pixel-wise for better controlling the objects in images. Geometry-guided adaptive instance normalization (GAIN) [33] modulates the activations using the texture while constraining the steepness of image gradients through a geometry-guided gate. This module can adaptively control the texture uniformity in different body parts for better person generating under the guidance of parsing mask. SEmantic-BridgeE (SEBE) [32] is a delicate and plug-n-play attention mechanism, which bridges the semantic chasm between word features and image features by using semantic segmentation mask.

Image composition. The first end-to-end learning-based image harmonization approach is proposed by Tsai et al. [29], which effectively captures context as well as semantic knowledge and greatly improves the quality of image composition. GP-GAN [34] leverages the strengths of the classical gradient-based approaches and GAN-based approaches to solve high-resolution image composition problems. GCC-GANs [5] adjusts the geometric and color consistency of the composited image firstly, and then polishes the boundary. In this way, objects of different shapes can be automatically combined with the background effortlessly. DoveNet [7] introduces the concept of domain verification into image composition and improves the compositing ef-
toward better synthesis, we design UniCoRN as a unified framework (Sec. 3.1). Furthermore, two core components are proposed: (i) Condition injection (Sec. 3.2), which is composed of cross-modality condition fusion module (CMCFM) and feature adaptive batch normalization (FABN). CMCFM is utilized for fusing background with other input conditions to control the repainted content generation, and then the fused feature modulates the normalized activations in FABN. (ii) Condition constraint (Sec. 3.3), named hierarchical cross-modality similarity model (HCMSM), which captures feature patterns at different levels and bridge the gap between the synthetic image and the color condition.

3. Methodology

To make this paper self-contained, we first review the repainting task. In previous works [32,33], it can be formulated in a two-phase manner: generating repainted regions under several conditional constraints and then compositing them with a provided background image seamlessly.

In the generation phase, given texture $z$, color description $x^c$, and geometry condition $x^g$, the generator $F^G$ synthesizes a raw repainted image $\hat{y}_r$, as shown in Eq. (1):

$$\hat{y}_r = F^G(z, x^c, x^g).$$ (1)

In the compositing phase, the compositing model $F^C$ estimates color tone parameters $(\rho, \tau)$ based on $\hat{y}_r$ and a desirable background $y^b$. Then the color tone of $\hat{y}_r$ can be adjusted towards $y^b$ through an affine transformation with $(\rho, \tau)$. The adjusted repainted image is denoted as $y^r$. As shown in Eq. (2) and Eq. (3):

$$(\rho, \tau) = F^C(\hat{y}_r, y^b),$$ (2)

$$y^r = \tanh(\rho \odot \hat{y}_r \oplus \tau),$$ (3)

where $\odot$ and $\oplus$ are element-wise multiplication and addition respectively. Finally, the model combines the adjusted repainted image with the input background under the mask $M$ to synthesize the complete image $y$, which can be formulated in Eq. (4). Here the mask value is 0 for background pixels and 1 for elsewhere.

$$y = M \odot y^r + (1 - M) \odot y^b.$$ (4)

However, existing methods assume dependency between phases and cause color-image incongruity. To take a step toward better synthesis, we design UniCoRN as a unified framework (Sec. 3.1). Furthermore, two core components are proposed: (i) Condition injection (Sec. 3.2), which is composed of cross-modality condition fusion module (CMCFM) and feature adaptive batch normalization (FABN). CMCFM is utilized for fusing background with other input conditions to control the repainted content generation, and then the fused feature modulates the normalized activations in FABN. (ii) Condition constraint (Sec. 3.3), named hierarchical cross-modality similarity model (HCMSM), which captures feature patterns at different levels and bridge the gap between the synthetic image and the color condition.

3.1. Framework

We propose a unified framework to relax the dependency of two phases in the CIR task. The one-step process of UniCoRN can be formulated as Eq. (5):

$$y^r = F^G(z, x^c, x^g, y^b).$$ (5)

The input conditions to $F^G$ are exactly the same as the existing method [33], including: (i) $z \sim \mathcal{N}(0, 1)$ denotes Gaussian noise vector for synthesizing diverse results. (ii) $x^c \in \mathbb{R}^{N_c \times N_c}$ is the multi-hot attribute, where $L \in \{0, 1\}$, $N_c$ denotes the number of attributes (e.g., coat color), and $N_v$ denotes the number of values (e.g., blue). (iii) $x^g \in \mathbb{R}^{N_g \times H \times W}$ is the segmentation mask. (iv) $y^b \in \mathbb{R}^{3 \times H \times W}$ represents the provided background image.

We build our generator $F^G$ based on GauGAN [22], which contains a series of FABN modules and convolutional layers. See Fig. 2 for an overview.

As shown in Fig. 2, the user-given conditions are injected into UniCoRN at the beginning of the network and in the middle of FABN. We broadcast the embedded color attribute under the guidance of geometry $x^g$ to make the color condition spatially-specific, denoted as $e^{x^g}$. In this way, the hidden feature $h$ contains both semantic and spatial information as the initial input to the generator. After a series of FABN and convolutional layers, $h$ is updated to enrich image details under the guidance of texture $z$, geometry $x^g$, and background $y^b$. Specifically, $h$ is refined in FABN and input conditions are fused in CMCFM.

As the condition constraint module, HCMSM is adopted in two ways: (i) The multi-grained attentive similarity loss proposed in HCMSM provides the supervision signals for whether the synthesized image is aligned with the input color condition. (ii) The label encoder in Fig. 2 is pretrained in HCMSM (similar to CMSM [36]) to guarantee the meaningfulness of the attribute embeddings.

There are three discriminators used in our model: (i) a joint-conditional-unconditional patch discriminator [16] to judge condition consistency and indicate the realness of each patch, (ii) a three-layer convolutional neural network
Figure 2. Overview of the proposed generator $F^{G}$. The color condition is first embedded and broadcast under the guidance of $x^g$, denoted as $e^{g_0}$. Then we convolve it and obtain the hidden feature $h$ as the initial input to $F^{G}$. Other input conditions are fed into FABN in the middle of $F^{G}$. The generator is composed of several FABN and convolutional layers. In FABN, conditions $x^g$, $y^b$, and $h$ are fused in CMCFM and then convolved to produce appearance parameters $\beta^p$ and $\gamma^a$, along with pattern parameters $\beta^b$ and $\gamma^b$ from texture $z$. The produced parameters are used to modulate $h$ after batch normalization. In CMCFM, geometry $x^g$ is convolved as a gate to fuse the hidden feature $h$ and the background feature.

3.2. Condition Injection

It is not feasible to design an individual injection module for each condition, because the interaction and dependency between conditions should be taken into account, e.g., geometry condition guides the spatial distribution of color condition and separates repainted regions from the background. Simply stacking all conditions together is not feasible, either, because conditions belong to different modalities and they are represented in different formats, e.g., geometry condition is segmentation mask while color condition is a set of vectors.

CMCFM and FABN are designed to solve the problems discussed above. CMCFM maps the background $y^b$ into a common high-dimensional feature space first, and then takes geometry $x^g$ as the gate to fuse the repainted feature and background feature spatially, as shown in Fig. 2.

Once the conditions are fused into a spatially-specific feature, they are injected into the FABN to produce appearance parameters $\beta^p$ and $\gamma^a$, along with pattern parameters $\beta^b$ and $\gamma^b$ from the texture condition. After summing separately, the produced spatially-adaptive parameters are multiplied and added to the normalized activation element-wise. See Fig. 2 for details. FABN is modified from GAIN [33] by removing the sigmoid function and modifying the multiplication to addition. This is because the fused feature after CMCFM is more complex and abstract instead of the pure geometry feature in GAIN, which makes it no longer suitable as the gate of texture pattern.

3.3. Condition Constraint

HCMSM consists of two pyramid-shaped encoders and a hierarchical attentive similarity model. Comparing to previous CMSM [36], the hierarchical structure can better constrain the color-image incongruity benefiting from its continuous feature space.

Image encoder. As shown in Fig. 3(a), we first divide the synthetic image into $n$ class regions, denoted as $(y'_0, y'_1, \ldots, y'_{n-1})$ under the guidance of geometry $x^g$. Then we feed them into the $n$-Group convolutional network
(G0, G1, . . . , Gn−1), where each group convolution corresponds to a class region. We extract m intermediate layer features for excavating multi-grained semantic information, represented as y′i,j ∈ RQ×H×W, where Q is the feature dimension, i ∈ {0, . . . , m − 1} and j ∈ {0, . . . , n − 1}.

Label encoder. As shown in Fig. 3(b), we introduce the attribute encoder in MISC [33] as our encoder unit Enc. As shown in Fig. 3(b), we introduce the hierarchical structure of encoders, we improve the similar-

ty between the encoders, we improve the similarity based on AttnGAN [36]. Specifically, given the color condition E and the synthetic image Y, we denote the color-image pair {Ei,t, Yi,t} as the i-th sample in a batch at the t-th feature level. Then we calculate the posterior probability of color Ei,t given image Yi,t using AttnGAN [36], as P(Ei,t|Yi,t). Finally, our multi-grained attentive similarity loss can be computed as:

$$L_m = -\sum_i \sum_t \log \left( P(E_i,t|Y_i,t)P(Y_i,t|E_i,t) \right). \quad (6)$$

3.4. Learning

To define the generative loss Lg, we introduce the discriminator first. We denote D as the largest one of the joint-conditioned unconditional patch discriminators proposed by Obj-GAN [16], which contains two parts: unconditional patch discriminator D0 and conditional patch discriminator D0. The process of discriminator prediction is written as p[y] = D0(y) and p[c|y] = D0(y, c), where y is the concatenation of y and y, c is the spatially-specific attribute after broadcasting, and p = {p1, . . . , pn} means a series of probabilities of patch realness, where Npix is the number of patches in discriminators. We define the generative loss Lg as:

$$L_g(F^G, D^I) = -\sum_{i=1}^{N_{pix}} (\lambda^u \log p^u_i[y] + \log p^I_i[y, e^C]). \quad (7)$$

Considering that the visual content should be seamlessly composited with background regions, we take a three-layer convolutional neural network Dc to separate repainted regions from the synthetic image following GCC-GAN [5], formulated as p = Dc(y). p = {p1, . . . , pNpix} indicates the probability that pixels are recognized as the painted ones, and Npix denotes the number of repainted pixels. We define the classifying loss as:

$$\mathcal{L}_c(F^G, D^C) = \frac{1}{N_{pix}} \sum_{i=1}^{N_{pix}} \log(1 - p^I_i). \quad (8)$$

An L1 loss is used in background regions to ensure the meaningfulness of background features:

$$\mathcal{L}_b(F^G) = \frac{1}{N_{pix}} \sum_{i=1}^{N_{pix}} \|y_i^b - \hat{y}_i^b\|_1, \quad (9)$$

where Npix is the number of background pixels.

The feature matching loss and perceptual loss are widely used to improve the synthetic quality of images. In our context, the feature matching loss [31] calculates the mean L1 distance of feature pairs extracted by discriminator DFM, defined as:

$$\mathcal{L}_{FM}(F^G, DFM) = \sum_{i=1}^{T_{FM}} \|D_i(y) - D_i(y^b)\|_1, \quad (10)$$

and the perceptual loss [9] takes a well-pretrained base network φ as an encoder to reduce the gap between image features, written as:

$$\mathcal{L}_p(F^G) = \sum_{i=1}^{T_p} \frac{1}{C_i H_i W_i} \|\phi(y) - \phi(y^b)\|_2^2, \quad (11)$$

where TFM andTp are the number of layers in DFM and φ, respectively.

Finally, we train our model with the full objective loss as:

$$\min_{F^G, D^I, D^C, DFM} \max_{\lambda, \lambda'} \mathcal{L}_g(F^G, D^I) + \lambda_c \mathcal{L}_c(D^C, F^G) + \lambda_b \mathcal{L}_b(F^G) + \lambda_{FM} \mathcal{L}_{FM}(F^G, DFM) + \lambda_p \mathcal{L}_p(F^G) + \lambda_m \mathcal{L}_m(F^G), \quad (12)$$

where Lm is the multi-grained attentive similarity loss to bridge the semantic gap between color conditions and synthetic images, formulated in Eq. (6).

Based on experiments using a held-out validation set, we set the hyperparameters as λc = 0.03, λb = 1.0, λFM = 10.0, λp = 10.0, and λm = 2.0.
Following the previous Quantitative evaluation metrics. We use Fréchet inception distance (FID), training images and 6K test images in this dataset and we resize them into 512 × 512 and resize them into 256 × 256 resolutions. Secondly, in order to avoid expensive manual annotation, we use the pretrained segmentation network DeepLabV2 [6] to compute the scene parsing mask for each image. After that, we pick out 7 major categories of objects for repainting, that is, cloud, flower, grass, hill, mountain, sky, and tree. Finally we divide the HSV color space into 10 intervals and count the proportion of pixels falling into each interval for every repainted object, and then all the interval proportions are concatenated into a vector as a color attribute. Therefore, at most 70 attributes are annotated in one image. We name this dataset LANDSCAPE-CIR.

For a fair comparison with previous methods, we also conduct experiments on VIP person parsing dataset [45]. This dataset provides RGB images and parsing masks, and we process them in the same way as MISC [33], which crops the images to keep one major person in each image and resizes them into 512 × 256 resolution. There are 42K training images and 6K test images in this dataset and we name it PERSON-CIR.

5. Experiments

Quantitative evaluation metrics. Following the previous works [32, 33], we use Fréchet inception distance (FID), R-precision, and M-score for performance evaluation. FID [12] is commonly used to measure the synthetic quality of images. R-precision [36] is used to evaluate whether generated images are well conditioned on the given color inputs. We mix 5 randomly sampled images with the synthetic image to calculate the accuracy of the color-image retrieval, using the same configuration as MISC [33]. The M-score [28] is used to measure the authenticity of images based on the detection model [44]. We randomly feed 100 synthetic images into the detection model to score every image. The lower the M-score, the more realistic the image is.

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

We quantitatively and qualitatively compare our UniCoRN with MISC [33], Weng et al. [32], Pavllo et al. [24], and SEAN [46] on PERSON-CIR and LANDSCAPE-CIR datasets. Note that except for MISC, other methods for comparison cannot be directly applied to the CIR task due to different input conditions. Thus we make some modifications and adapt them to the CIR task. We show synthetic images in Fig. 4 and Fig. 5, and evaluation scores in Tab. 1 to demonstrate that UniCoRN achieves better performance in synthetic quality, condition consistency, and compositing effect than other state-of-the-art methods.

MISC [33] is a conventional two-phase CIR model, which uses an additional compositing model to adjust the color tone of the repainted foreground for high robustness and training stability. However, due to the unstable gradient backpropagation caused by two-phase design, the synthetic quality is unsatisfactory with artifacts, e.g., the top sample in Fig. 4(b) appears silver spots artifacts in the red clothes.

Weng et al.’s method [32] tackles the CIR task by representing color condition in the language form instead of attribute. To unify the form of color condition, we replace its injection module (SEBE) with the attribute injection module (GAIN) proposed by MISC [33]. Similar to MISC [33], this two-phase CIR method also produces obvious artifacts, e.g., the black hole in the tree of the top sample in Fig. 5(c).

Pavllo et al.’s method [24] makes progress in synthesizing complex scenes with user-given attributes and masks. Considering that the background is user-specified, we simply utilize their foreground generator. However, due to the lack of color adjustment, the boundary between repainted foreground and background is obvious, e.g., the person outline of the top and bottom samples in Fig. 4(d).

SEAN [46] proposes a novel normalization block for GANs conditioned on style matrices extracted from input images and segmentation masks. Note that we cannot directly evaluate it on CIR task due to the different inputs. As a comparison, we replace FABN with the efficient SEAN block to fit it into our unified framework. Compared with FABN, the SEAN block takes more parameters and causes color-image mismatching, e.g., the grey clouds in the bot-

1License: CC-BY 4.0. URL: https://www.flickr.com.
bottom sample of Fig. 5(e) is inconsistent with the input “blue” condition. Thanks to our unified framework, it is slightly worse than UniCoRN quantitatively, as shown in Tab. 1.

5.2. User Study

We further conduct user study experiments on three datasets to evaluate whether our results are favored by human observers. We provide input conditions, original images, and candidates generated from five different methods: MISC [33], Weng et al. [32], Pavllo et al. [24], SEAN [46], and our UniCoRN. Participants are asked to choose the most visually pleasing result according to input conditions and the original image. The experiment on each dataset includes 100 sets of synthetic images randomly selected. We publish the experiments on Amazon Mechanical Turk (AMT), and each experiment is completed by 25 participants. As shown in Fig. 6, our UniCoRN performs better than other comparison methods, confirming its subjective advantages.

5.3. Ablation Study

The ablation study focuses on the effectiveness of pyramid-shaped encoders in HCMSM and the benefit of the unified network. The evaluation scores and synthetic images of the ablation study are shown in Tab. 1 and Fig. 7.

CMSM denotes we replace our HCMSM with the original CMSM proposed in AttnGAN [36], where the image encoder and label encoder only care about the global and high-level semantic information, thus more likely to cause
color-image incongruity, e.g., grey stripe artifacts in the white clothes presented in the top left sample in Fig. 7(b).

**SINGLE** reduces the number of layers from \( m \) to 1 to measure the necessity of pyramid-shaped encoders, which provide semantic information at different levels. As the number of layers decreases, the semantic information becomes inadequate, which brings uncertainty and lowers the synthetic quality. This ablation version fails in boundary harmonization, e.g., the bottom sample in Fig. 7(c).

**W/o assist** removes connections between encoder units in the label encoder and flattens the encoder structure. Losing the hierarchy, all the features are extracted in the same semantic level, which makes the color tone look artificial, e.g., the top right sample in Fig. 7(d).

**Two-phase** injects the background condition into an additional compositing model instead of CMCFM, similar to two-phase methods in previous works [32, 33]. Limited by the two-phase dependency, **Two-phase** can not reach high synthetic quality, e.g., the bottom sample in Fig. 7(e).

### 5.4. Controllability Study

We demonstrate the robustness of UniCoRN in synthesizing images by modified input conditions, e.g., color interpolation and geometry manipulation, shown in Fig. 8.

### 6. Conclusion

We propose a unified framework to solve the CIR task. Compared with the existing two-phase CIR methods, UniCoRN relaxes the two-phase dependency and introduces hierarchical structure into condition constraint, which reaches higher synthetic quality, better condition consistency, and more realistic compositing effect.

**Limitation.** Considering the necessity of constructing the interaction and dependency relationship between different cross-modality inputs, our model includes a large number of learnable parameters (103.1M). In future work, we will simplify our network into a lightweight structure with fewer parameters and deploy it on mobile devices.

### Acknowledgements

This work is supported by National Natural Science Foundation of China under Grand No. 62136001, 62088102, 61872012.
References


[36] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In CVPR, 2018. 2, 3, 4, 5, 6, 7

[37] Jie Yang, Zhiquan Qi, and Yong Shi. Learning to incorporate structure knowledge for image inpainting. In AAAI, 2020. 1


[40] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In ICCV, 2017. 2


[44] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning rich features for image manipulation detection. In CVPR, 2018. 6


7. Appendix

7.1. Unstable two-phase gradient backpropagation

We show the box plots of gradient magnitude in Fig. 9. The lateral axis denotes the percentage of the repainted area to the whole image area. The vertical axis denotes the gradient magnitude distribution of the last convolutional layer that generates visual contents.

The two-phase model discards the generated meaningless background, which truncates the gradient, making the gradient backpropagation unstable and the weights difficult to be optimized, as shown in Fig. 9 (top row). Noting that with two-phase model, the gradient magnitude distribution is related to the area of repainted regions. The gradient magnitude distribution of our unified model is shown in Fig. 9 (bottom row), demonstrating that the gradients of our model are irrelevant to the area of repainted regions. In this way, the gradient backpropagation is stable, which is one of the reasons why our model performs better than the two-phase models.

7.2. Supplemental dataset description

The LANDSCAPE-CIR dataset we created concentrates on landscape generation, containing 31K high-resolution images. Examples of images in the dataset are shown in Fig. 10. We use the pretrained segmentation model [1] to compute the scene parsing mask for each image, and then annotate color attributes by measuring the distribution of pixels on the HSV color intervals. The color ranges are defined in Tab. 2.

Since there are seven major categories of objects in the LANDSCAPE-CIR dataset, and every category contains ten color attributes, we set the values of $N_v$, $N_c$, $N_g$ (defined in the Sec. 3 of the main paper) as 10, 7, 7. In PERSON-CIR dataset, these values are consistent with the existing work [6], which are 10, 11, 19, respectively.

Among the seven categories in LANDSCAPE-CIR, “flower” is a purposely collected challenging category for testing the robustness of the geometry injection module. It is difficult for CIR models to repaint flowers because the off-the-shelf segmentation model cannot separate the flower and the background precisely, which makes the geometry information less accurate, as shown in Fig. 11(b). For existing CIR models (including UniCoRN), users need to correct the segmentation mask by hand to reset the repainted area. Despite carefully depicted masks are provided, the quality of the synthesized images still remains to be unsatisfactory as the inaccurate geometry has compromised the model during the training stage, as shown in Fig. 11(e). We hope this could also inspire the following research about CIR task.

7.3. Hyperparameters sensitivity experiments

In our method, all datasets share the same hyperparameters, and we show the sensitivity analysis of parameters on PERSON-CIR to demonstrate the stability of UniCoRN, as shown in Tab. 3. The hyperparameters listed here are formulated in the Sec. 3 of the main paper.

7.4. Additional quantitative and qualitative results

We qualitatively compare our UniCoRN with previous methods, e.g., MISC [6], Weng et al. [5], Pavllo et al. [3], and SEAN [8]. We also create four baselines to demonstrate the effectiveness of pyramid-shaped encoders in HCMSM and the benefit of the unified network, whose details are presented in the Sec. 5 of the main paper.

To draw a more comprehensive evaluation of UniCoRN and other comparison methods mentioned above, we additionally use three quantitative metrics to measure the synthetic image quality, i.e., inception score (IS) [4], LPIPS [7], and NIQE [2]. IS [4] is commonly used as an evaluation metric to measure the image quality for GANs. LPIPS [7] is a perceptual similarity metric to evaluate the distance between the generated images and original images with VGG backbone. NIQE [2] calculates the no-reference image quality score for generated images, where a smaller score indicates better perceptual quality. Quantitative comparison and ablation experiments with additional metrics are shown in Tab. 4.

We show more experimental comparison and ablation results on PERSON-CIR in Fig. 12 and Fig. 13. Comparison and ablation results on LANDSCAPE-CIR dataset are shown in Fig. 14 and Fig. 15.
Figure 9. The box plots of gradient magnitude with the area of repainted regions in two-phase method MISC [6] (top row) and in our UniCoRN (bottom row). Please zoom in for details.

Table 2. The range of color intervals in HSV color space. Some colors covering multiple intervals are denoted with numbers at subscripts.

<table>
<thead>
<tr>
<th>Color</th>
<th>Abbr.</th>
<th>H range</th>
<th>S range</th>
<th>V range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>BK</td>
<td>0-180</td>
<td>0-255</td>
<td>0-39</td>
</tr>
<tr>
<td>Gray0</td>
<td>GY</td>
<td>0-180</td>
<td>0-12</td>
<td>40-255</td>
</tr>
<tr>
<td>Gray1</td>
<td>GY</td>
<td>10-180</td>
<td>13-26</td>
<td>40-255</td>
</tr>
<tr>
<td>White</td>
<td>WT</td>
<td>0-180</td>
<td>0-12</td>
<td>40-255</td>
</tr>
<tr>
<td>Red0</td>
<td>RD</td>
<td>170-180</td>
<td>164-255</td>
<td>40-255</td>
</tr>
<tr>
<td>Red1</td>
<td>RD</td>
<td>0-7</td>
<td>164-255</td>
<td>40-255</td>
</tr>
<tr>
<td>Orange</td>
<td>OR</td>
<td>8-20</td>
<td>27-255</td>
<td>40-255</td>
</tr>
<tr>
<td>Yellow</td>
<td>YL</td>
<td>21-30</td>
<td>27-255</td>
<td>40-255</td>
</tr>
<tr>
<td>Green0</td>
<td>GR</td>
<td>31-78</td>
<td>27-255</td>
<td>40-255</td>
</tr>
<tr>
<td>Green1</td>
<td>GR</td>
<td>79-85</td>
<td>154-255</td>
<td>40-255</td>
</tr>
<tr>
<td>Blue0</td>
<td>BL</td>
<td>79-85</td>
<td>27-151</td>
<td>40-255</td>
</tr>
<tr>
<td>Blue1</td>
<td>BL</td>
<td>86-123</td>
<td>27-255</td>
<td>40-255</td>
</tr>
<tr>
<td>Purple0</td>
<td>PP</td>
<td>124-131</td>
<td>201-255</td>
<td>40-255</td>
</tr>
<tr>
<td>Purple1</td>
<td>PP</td>
<td>132-143</td>
<td>27-255</td>
<td>40-255</td>
</tr>
<tr>
<td>Pink0</td>
<td>PK</td>
<td>144-170</td>
<td>27-255</td>
<td>40-255</td>
</tr>
<tr>
<td>Pink1</td>
<td>PK</td>
<td>170-180</td>
<td>27-163</td>
<td>40-255</td>
</tr>
<tr>
<td>Pink2</td>
<td>PK</td>
<td>0-7</td>
<td>27-163</td>
<td>40-255</td>
</tr>
</tbody>
</table>

7.5. Additional controllability study

In this section, we demonstrate the controllability of UniCoRN in different application scenarios. By interpolating the input color and modifying the input geometry, the repainted regions can be greatly changed. The synthetic results on PERSON-CIR and LANDSCAPE-CIR are provided in Fig. 16 and Fig. 17, respectively.

7.6. Network architecture details

In this section, we present the details regarding the network architecture of UniCoRN. We show the architecture of conditional generators in UniCoRN for different resolutions in Fig. 18, and details of the image encoder and the label encoder are shown in Fig. 19. Please see the definitions of symbols in the Sec. 3 of the main paper.

References

Table 4. Quantitative comparison and ablation experiments with additional metrics. ↑ (↓) indicates larger (smaller) values are better. The best performances are highlighted in **bold**.

<table>
<thead>
<tr>
<th>Category</th>
<th>Methods</th>
<th>PERSON-CIR</th>
<th>LANDSCAPE-CIR (256 × 256)</th>
<th>LANDSCAPE-CIR (512 × 512)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>IS ↑</td>
<td>LPIPS ↓</td>
<td>NRI ↓</td>
</tr>
<tr>
<td><strong>Comparison</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MISC [6]</td>
<td>5.11 ± 0.14</td>
<td>0.1317</td>
<td>21.39</td>
</tr>
<tr>
<td></td>
<td>Weng et al. [5]</td>
<td>4.87 ± 0.12</td>
<td>0.1264</td>
<td>21.15</td>
</tr>
<tr>
<td></td>
<td>Pavlin et al. [3]</td>
<td>4.94 ± 0.23</td>
<td>0.1411</td>
<td>21.37</td>
</tr>
<tr>
<td></td>
<td>SEAN [8]</td>
<td>4.88 ± 0.17</td>
<td>0.1120</td>
<td>21.04</td>
</tr>
<tr>
<td><strong>Ablation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CMSM</td>
<td>4.90 ± 0.13</td>
<td>0.1152</td>
<td>21.32</td>
</tr>
<tr>
<td></td>
<td>SINGLE</td>
<td>4.91 ± 0.14</td>
<td>0.1151</td>
<td>21.33</td>
</tr>
<tr>
<td></td>
<td>W/o assist</td>
<td>4.94 ± 0.16</td>
<td>0.1068</td>
<td>21.42</td>
</tr>
<tr>
<td></td>
<td>Two-phase</td>
<td>4.92 ± 0.16</td>
<td>0.1144</td>
<td>21.42</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>UniCoRN</td>
<td><strong>5.23 ± 0.19</strong></td>
<td><strong>0.1032</strong></td>
<td><strong>20.60</strong></td>
</tr>
</tbody>
</table>

Figure 10. Image examples in LANDSCAPE-CIR. There are seven major categories picked up to be repainted, that is, cloud, flower, grass, hill, mountain, sky, and tree.

(a) (b) (c) (d) (e)

Figure 11. Challenging examples in LANDSCAPE-CIR. (a) Original image. (b) The inaccurate segmentation result produced by the off-the-shelf segmentation model DeepLabV2 [1]. (c) Visualization of the inaccurate segmentation mask in RGB image. (d) Correct the segmentation mask by hand. (e) Repaint the image with the corrected mask, whose quality needs to be improved.
Figure 12. Qualitative comparison with several state-of-the-art methods on PERSON-CIR. The green dotted boxes highlight the artifacts. (a) Input conditions: color (left), geometry (top right), and background (bottom right); texture is omitted here. (b) MISC [6]. (c) Weng et al. [5]. (d) Pavllo et al. [3]. (e) SEAN [8]. (f) Our results. (g) Original images. Note that we repaint all samples with the original conditions (the same as the original images). Please zoom in for details and observe repainted regions according to input conditions.
Figure 13. Ablation study with different variants of the proposed method on PERSON-CIR. The green dotted boxes highlight the artifacts. (a) Input conditions: color (left), geometry (top right), and background (bottom right); texture is omitted here. (b) CMSM. (c) SINGLE. (d) W/o assist. (e) Two-phase. (f) Our results. (g) Original images. Note that we repaint all samples with the original conditions (the same as the original images). Please zoom in for details and observe repainted regions according to input conditions.
Figure 14. Qualitative comparison with several state-of-the-art methods on LANDSCAPE-CIR. The green dotted boxes highlight the artifacts. (a) Input conditions: color (left), geometry (top right), and background (bottom right); texture is omitted here. (b) MISC [6]. (c) Weng et al. [5]. (d) Pavllo et al. [3]. (e) SEAN [8]. (f) Our results. (g) Original images. Note that we repaint all samples with the original conditions (the same as the original images). Please zoom in for details and observe repainted regions according to input conditions.
Figure 15. Ablation study with different variants of the proposed method on LANDSCAPE-CIR. The green dotted boxes highlight the artifacts. (a) Input conditions: color (left), geometry (top right), and background (bottom right); texture is omitted here. (b) CMSM. (c) SINGLE. (d) W/o assist. (e) Two-phase. (f) Our results. (g) Original images. Note that we repaint all samples with the original conditions (the same as the original images). Please zoom in for details and observe repainted regions according to input conditions.
Figure 16. Controllability of color and geometry conditions on PERSON-CIR. Please zoom in for details.

Figure 17. Controllability of color and geometry conditions on LANDSCAPE-CIR. Please zoom in for details.
Figure 18. The architecture of conditional generators in UniCoRN. (a) The generator for $256 \times 256$ resolution. (b) The generator for $512 \times 512$ resolution. $h$, $x^g$, and $y^b$ denote the input hidden feature, the input geometry condition, and the input background condition, respectively.
Figure 19. The architecture of encoders. (a) The image encoder. (b) The label encoder. (c) The unit of the label encoder ($\text{Enc}_i^c$).