UniCoRN: A Unified Conditional Image Repainting Network

Jimeng Sun[#1] Shuchen Weng[#2] Zheng Chang[1] Si Li[*1] Boxin Shi[2,3,4]
[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[3]Institute for Artificial Intelligence, Peking University
[4]Beijing Academy of Artificial Intelligence
{sjm,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 **Uni**fied **Co**nditional image **R**epainting **N**etwork (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.

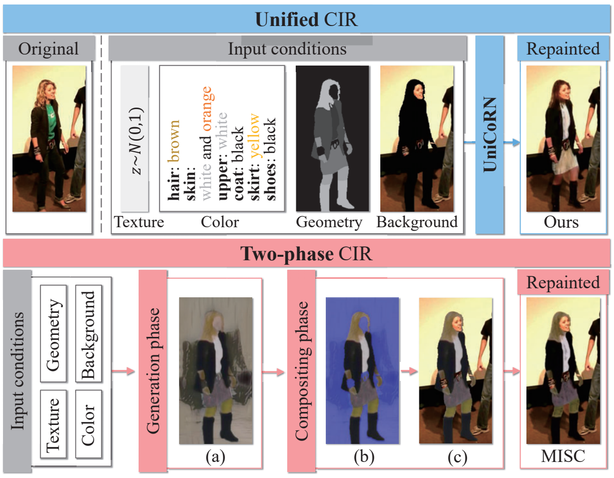


Figure 1. Illustration of the CIR task. Given an original image and input conditions, the model needs to repaint the user-specified regions (geometry) to synthesize a new appearance, and composite it with the input background regions seamlessly. The bottom row visualizes the pipeline of a representative two-phase CIR method called "MISC" [33], where (a), (b), and (c) are the output of the generation phase, visualization of meaningless background with purple mask, and the input of the compositing model, respectively. Compared with the two-phase CIR repainted result (highlighted in pink), our unified CIR repainted result (highlighted in blue) improves the appearance of the repainted image in synthetic quality, condition consistency, and compositing effect.

Existing CIR methods [32,33] are implemented in a two-

[#] Equal contributions. [*] Corresponding author.

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 **Uni**fied **Co**nditional image **R**epainting **N**etwork, 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-BridegE (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 effect. The composition model in MISC [33] enables the bounding mechanism and the spatial adaptability to reduce the risk of the gradient vanishing pitfall. The novel piecewise value function [32] aims to break through the latent ceiling of fidelity in content compositing.

Conditional image repainting. The CIR task is first formulated by Weng *et al.* [32] as an advanced image editing technique: the model is trained to repaint the visual content in a specified image conditioned on user inputs. Specifically, the user inputs should cover at least three aspects, *e.g.*, geometry, color, and texture. The color condition can be expressed by attribute or language, corresponding to different application scenarios. The conditional person image synthesis [33] is also a kind of CIR task, where the repainted visual content is limited to the person with clear region division, thus they choose attribute as color condition for simplicity and efficiency.

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 $\hat{y}^{\mathrm{r}}$, as shown in Eq. (1):

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

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

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

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

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.

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

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):

$$y^{\mathrm{r}} = F^{\mathrm{G}}(z, x^{\mathrm{c}}, x^{\mathrm{g}}, y^{\mathrm{b}}). \tag{5}$$

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)* $x^{\mathrm{c}} \in \mathbb{L}^{N_{\mathrm{c}} \times N_{\mathrm{v}}}$ is the multi-hot attribute, where $\mathbb{L} \in \{0, 1\}$, N_{c} denotes the number of attributes (*e.g.*, coat color), and N_{v} denotes the number of values (*e.g.*, blue). *(iii)* $x^{\mathrm{g}} \in \mathbb{L}^{N_{\mathrm{g}} \times H \times W}$ is the segmentation mask, where $\mathbb{L} \in \{0, 1\}$, and N_{g}, H, and W denote the number of parts in repainted regions, image height, and width, respectively. *(iv)* $y^{\mathrm{b}} \in \mathbb{R}^{3 \times H \times W}$ 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^{gc}. 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

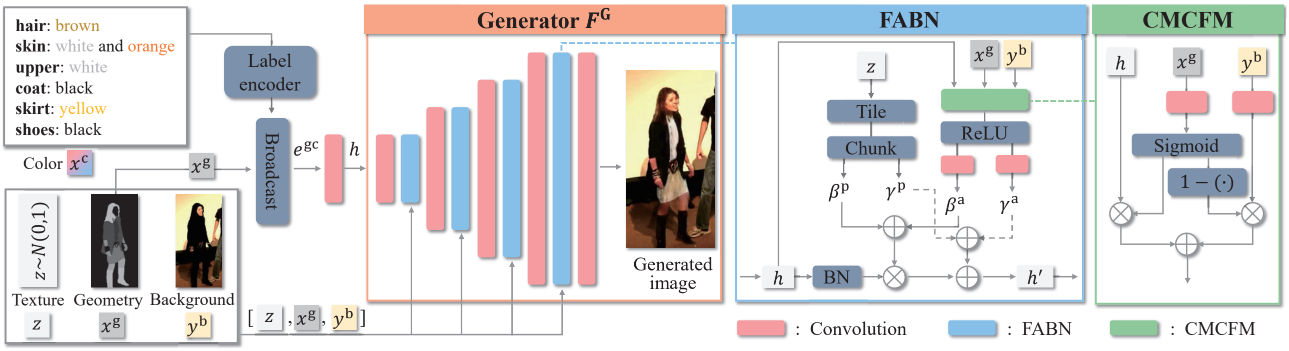


Figure 2. Overview of the proposed generator F^{G}. The color condition is first embedded and broadcast under the guidance of x^{g}, denoted as e^{gc}. Then we convolve it and obtain the hidden feature h as the initial input to F^{G}. Other input conditions are fed into FABN in the middle of F^{G}. The 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 β^{a} and γ^{a}, along with pattern parameters β^{p} and γ^{p} 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.

to supervise the color tone harmonious degree between repainted regions and the background, and *(iii)* a multi-scale discriminator [31] for calculating feature matching loss to distinguish real images and synthetic images at different feature levels.

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 provided 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 β^{a} and γ^{a}, along with pattern parameters β^{p} and γ^{p} 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.

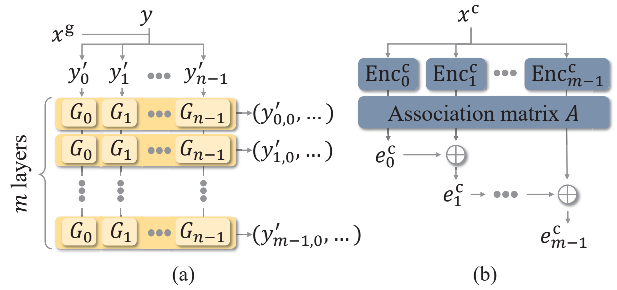


Figure 3. Two pyramid-shaped encoders in HCMSM. (a) The image encoder is an n-Group convolutional network $(G_0, G_1, \dots, G_{n-1})$, and extracts m features from m intermediate layers as outputs. (b) The label encoder consists of m encoder units $\mathrm{Enc}^{\mathrm{c}}$ and represents attributes in different semantic levels. With element-wise addition, units are connected to form a hierarchical structure.

Note that the texture condition z is only fed into FABN instead of fusing with other conditions in CMCFM. That is because the texture condition is Gaussian noise and needs to be highlighted to make generated results more diverse.

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 $(y'_0, y'_1, \dots, y'_{n-1})$ under the guidance of geometry x^{g}. Then we feed them into the n-Group convolutional network

$(G_0, G_1, \dots, G_{n-1})$, where each group convolution corresponds to a class region. We extract m intermediate layer features for excavating multi-grained semantic information, represented as $y'_{i,j} \in \mathbb{R}^{Q \times H \times W}$, where Q is the feature dimension, $i \in \{0, \dots, m-1\}$ and $j \in \{0, \dots, n-1\}$.

Label encoder. As shown in Fig. 3(b), we introduce the attribute encoder in MISC [33] as our encoder unit $\mathrm{Enc}^{\mathrm{c}}$, and we connect m units to form a pyramid-like structure as our label encoder to extract semantic features at different levels, inspired by pyramid methods [1, 3]. In detail, given color attribute x^{c}, each unit encodes it separately first, written as $\hat{e}^{\mathrm{c}}_i = \mathrm{Enc}^{\mathrm{c}}_i(x^{\mathrm{c}})$, where $\hat{e}^{\mathrm{c}}_i \in \mathbb{R}^{Q \times N_{\mathrm{c}}}$, Q is the embedding dimension, and $i \in \{0, \dots, m-1\}$. Then, specifying the association matrix between attributes and image class regions as $A \in \mathbb{L}^{N_{\mathrm{c}} \times N_{\mathrm{g}}}$ s.t. $\forall\ i, j\ A[i, j] \geq 0$, $\forall\ j\ \sum_i A[i, j] = 1$, we remap the embedded attribute into image class regions $e^{\mathrm{c}}_i = \hat{e}^{\mathrm{c}}_i A$, where $e^{\mathrm{c}}_i \in \mathbb{R}^{Q \times N_{\mathrm{g}}}$, N_{c} and N_{g} denote the number of attributes and parts in repainted regions respectively. Finally, we get the $(i+1)$-th level feature e^{c}_{i+1} by vector addition as $e^{\mathrm{c}}_{i+1} = e^{\mathrm{c}}_i + \mathrm{Enc}^{\mathrm{c}}_{i+1}(x^{\mathrm{c}})A$ and the first level feature as $e^{\mathrm{c}}_0 = \mathrm{Enc}^{\mathrm{c}}_0(x^{\mathrm{c}})A$.

Attentive similarity model. To take advantage of the hierarchical structure of encoders, we improve the similarity module based on AttnGAN [36]. Specifically, given the color condition E and the synthetic image Y, we denote the color-image pair $\{E_{i,t}, Y_{i,t}\}$ as the t-th sample in a batch at the i-th feature level. Then we calculate the posterior probability of color $E_{i,t}$ being matching with image $Y_{i,t}$ following AttnGAN [36], as $P(E_{i,t}|Y_{i,t})$. Finally, our multi-grained attentive similarity loss can be computed as:

$$\mathcal{L}_{\mathrm{m}} = -\sum_{i}^{m}\sum_{t}^{T} \log\Big(P(E_{i,t}|Y_{i,t})P(Y_{i,t}|E_{i,t})\Big). \tag{6}$$

3.4. Learning

To define the generative loss $\mathcal{L}_{\mathrm{g}}$, we introduce the discriminator first. We denote D^{I} as the largest one of the joint-conditional-unconditional patch discriminators proposed by Obj-GAN [16], which contains two parts: unconditional patch discriminator $D^{\mathrm{I}}_{\mathrm{u}}$ and conditional patch discriminator $D^{\mathrm{I}}_{\mathrm{c}}$. The process of discriminator prediction is written as $\mathbf{p}^{\mathrm{u}}[\bar{y}] = D^{\mathrm{I}}_{\mathrm{u}}(\bar{y})$ and $\mathbf{p}^{\mathrm{c}}[\bar{y}, e^{\mathrm{gc}}] = D^{\mathrm{I}}_{\mathrm{c}}(\bar{y}, e^{\mathrm{gc}})$, where $\bar{y}$ is the concatenation of y and y^{r}, e^{gc} is the spatially-specific attribute after broadcasting, and $\mathbf{p} = \{p_1, \dots, p_i, \dots, p_{N^{\mathrm{pat}}}\}$ means a series of probabilities of patch realness, where N^{pat} is the number of patches in discriminators. We define the generative loss $\mathcal{L}_{\mathrm{g}}$ as:

$$\mathcal{L}_{\mathrm{g}}(F^{\mathrm{G}}, D^{\mathrm{I}}) = -\sum_{i=1}^{N^{\mathrm{pat}}} (\lambda^{\mathrm{u}} \log p^{\mathrm{u}}_i[\bar{y}] + \log p^{\mathrm{c}}_i[\bar{y}, e^{\mathrm{gc}}]). \tag{7}$$

Considering that the visual content should be seamlessly composited with background regions, we take a three-layer convolutional neural network D^{C} to separate repainted regions from the synthetic image following GCC-GAN [5], formulated as $\mathbf{p}^{\mathrm{r}} = D^{\mathrm{C}}(y)$. $\mathbf{p}^{\mathrm{r}} = \{p^{\mathrm{r}}_1, \dots, p^{\mathrm{r}}_i, \dots, p^{\mathrm{r}}_{N^{\mathrm{r}}_{\mathrm{pix}}}\}$ indicates the probability that pixels are recognized as the repainted ones, and $N^{\mathrm{r}}_{\mathrm{pix}}$ denotes the number of repainted pixels. We define the compositing loss as:

$$\mathcal{L}_{\mathrm{c}}(F^{\mathrm{G}}, D^{\mathrm{C}}) = \frac{1}{N^{\mathrm{r}}_{\mathrm{pix}}} \sum_{i=1}^{N^{\mathrm{r}}_{\mathrm{pix}}} \log(1 - p^{\mathrm{r}}_i). \tag{8}$$

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

$$\mathcal{L}_{\mathrm{b}}(F^{\mathrm{G}}) = \frac{1}{N^{\mathrm{b}}_{\mathrm{pix}}} \sum_{i=1}^{N^{\mathrm{b}}_{\mathrm{pix}}} \left\| y^{\mathrm{r}}_i - y^{\mathrm{b}}_i \right\|_1, \tag{9}$$

where $N^{\mathrm{b}}_{\mathrm{pix}}$ 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 D^{FM}, defined as:

$$\mathcal{L}_{\mathrm{FM}}(F^{\mathrm{G}}, D^{\mathrm{FM}}) = \sum_{i=1}^{T_{\mathrm{FM}}} \left\| D_i(y) - D_i(y^{\mathrm{b}}) \right\|_1, \tag{10}$$

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

$$\mathcal{L}_{\mathrm{p}}(F^{\mathrm{G}}) = \sum_{i}^{T_{\mathrm{p}}} \frac{1}{C_i H_i W_i} \left\| \phi(y) - \phi(y^{\mathrm{b}}) \right\|_2^2, \tag{11}$$

where T_{FM} and T_{p} are the number of layers in D^{FM} and ϕ, respectively.

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

$$\min_{F^{\mathrm{G}}} \max_{D^{\mathrm{I}}, D^{\mathrm{C}}, D^{\mathrm{FM}}} \mathcal{L}_{\mathrm{g}}(D^{\mathrm{I}}, F^{\mathrm{G}}) + \lambda_{\mathrm{c}}\mathcal{L}_{\mathrm{c}}(D^{\mathrm{C}}, F^{\mathrm{G}}) + \lambda_b \mathcal{L}_{\mathrm{b}}(F^{\mathrm{G}}) + \lambda_{\mathrm{FM}}\mathcal{L}_{\mathrm{FM}}(F^{\mathrm{G}}, D^{\mathrm{FM}}) + \lambda_{\mathrm{p}}\mathcal{L}_{\mathrm{p}}(F^{\mathrm{G}}) + \lambda_{\mathrm{m}}\mathcal{L}_{\mathrm{m}}(F^{\mathrm{G}}), \tag{12}$$

where $\mathcal{L}_m$ 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 $\lambda_{\mathrm{c}} = 0.03$, $\lambda_{\mathrm{b}} = 1.0$, $\lambda_{\mathrm{FM}} = 10.0$, $\lambda_{\mathrm{p}} = 10.0$, and $\lambda_{\mathrm{m}} = 2.0$.

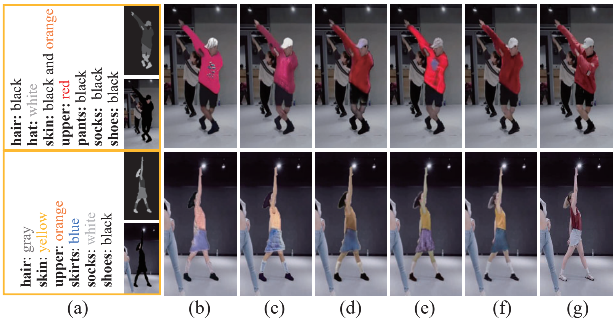


Figure 4. Qualitative comparison with state-of-the-art methods on PERSON-CIR. (a) Input conditions: color (left), geometry (top right), and background (bottom right); texture is omitted here. (b) MISC [33]. (c) Weng *et al.* [32]. (d) Pavllo *et al.* [24]. (e) SEAN [46]. (f) Our results. (g) Original images. Note that we repaint the top sample with the original conditions (the same as the original image), and the bottom sample with the edited conditions. Please zoom in for details.

4. Datasets

Existing CIR methods are mainly evaluated on synthetic persons [45], birds [30], and stuff [4]. To further expand the application scenarios of the CIR task, we create a dataset which concentrates on landscape generation. First of all, we download 31K high-resolution images from Flickr[1] including 28K training images and 3K test images, then we resize them into 256 $\times$ 256 and 512 $\times$ 512 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 $\times$ 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-

[1] License: CC-BY 4.0. URL: https://www.flickr.com.

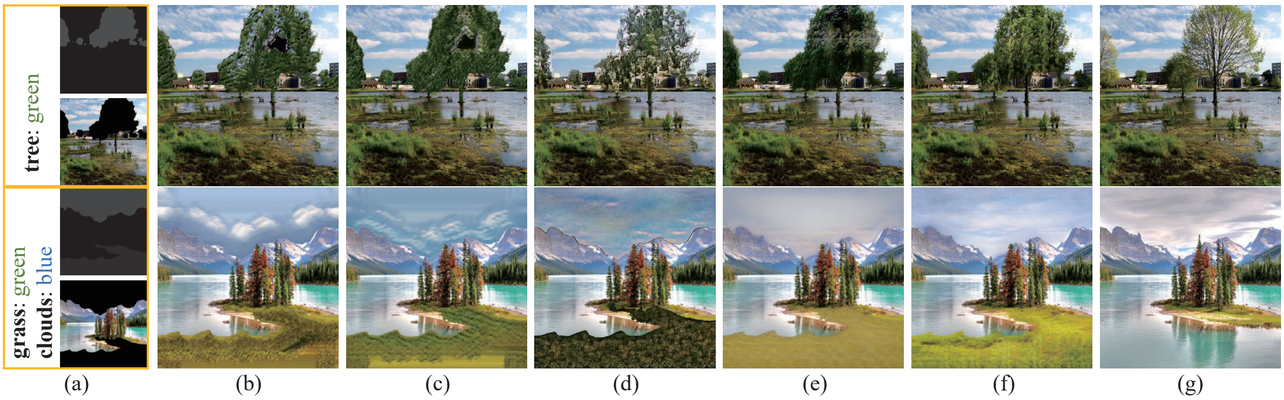


Figure 5. Qualitative comparison with state-of-the-art methods on LANDSCAPE-CIR. (a) Input conditions: color (left), geometry (top right), and background (bottom right); texture is omitted here. (b) MISC [33]. (c) Weng *et al.* [32]. (d) Pavllo *et al.* [24]. (e) SEAN [46]. (f) Our results. (g) Original images. Note that we repaint the top sample with the original conditions (the same as the original image), and the bottom sample with the edited conditions. Please zoom in for details.

Table 1. Quantitative experiments include comparison and ablation. ↑ (↓) indicates larger (smaller) values are better. The best performances are highlighted in **bold**.

Category	Methods	PERSON-CIR			LANDSCPAE-CIR (256 × 256)			LANDSCPAE-CIR (512 × 512)		
		FID ↓	R-prcn (%) ↑	M-score ↓	FID ↓	R-prcn (%) ↑	M-score ↓	FID ↓	R-prcn (%) ↑	M-score ↓
Comparison	MISC [33]	16.09	93.59	3.86	18.12	95.17	7.45	26.97	94.15	54.96
	Weng *et al.* [32]	15.59	93.02	3.89	18.10	96.51	7.27	25.17	93.33	54.64
	Pavllo *et al.* [24]	18.82	85.38	19.20	17.70	87.71	50.49	21.47	84.59	73.65
	SEAN [46]	13.74	96.11	4.62	14.71	96.73	3.18	20.00	96.96	15.66
Ablation	CMSM	15.03	93.02	6.01	13.35	92.78	3.24	20.13	92.11	22.35
	SINGLE	12.40	96.87	3.72	13.30	93.88	3.75	19.61	94.15	16.38
	W/o assist	12.21	97.23	5.76	12.71	97.01	3.23	21.42	94.37	15.11
	Two-phase	16.03	96.98	4.78	18.94	97.41	5.87	22.64	97.04	20.32
Ours	UniCoRN	**11.45**	**97.42**	**3.56**	**9.96**	**97.74**	**3.14**	**18.63**	**97.33**	**14.42**

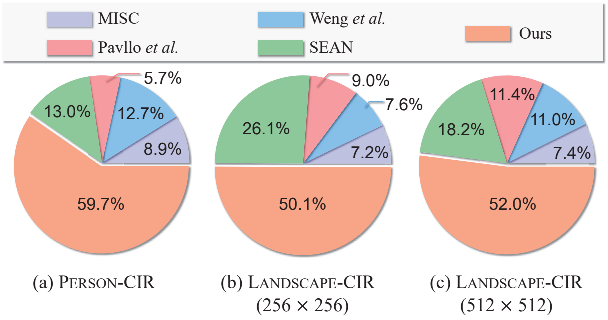
tom sample of Fig. 5(e) is inconsistent with the input "blue" condition. Thanks to our unified framework, it is slightly worse than UniCoRN quantitatively, as shown in Tab. 1.

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.



(a) PERSON-CIR (b) LANDSCAPE-CIR (256 × 256) (c) LANDSCAPE-CIR (512 × 512)

Figure 6. User study results on (a) PERSON-CIR, (b) LANDSCAPE-CIR (256×256), and (c) LANDSCAPE-CIR (512×512) datasets. Our UniCoRN achieves obviously higher scores on three datasets than other comparison methods.

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

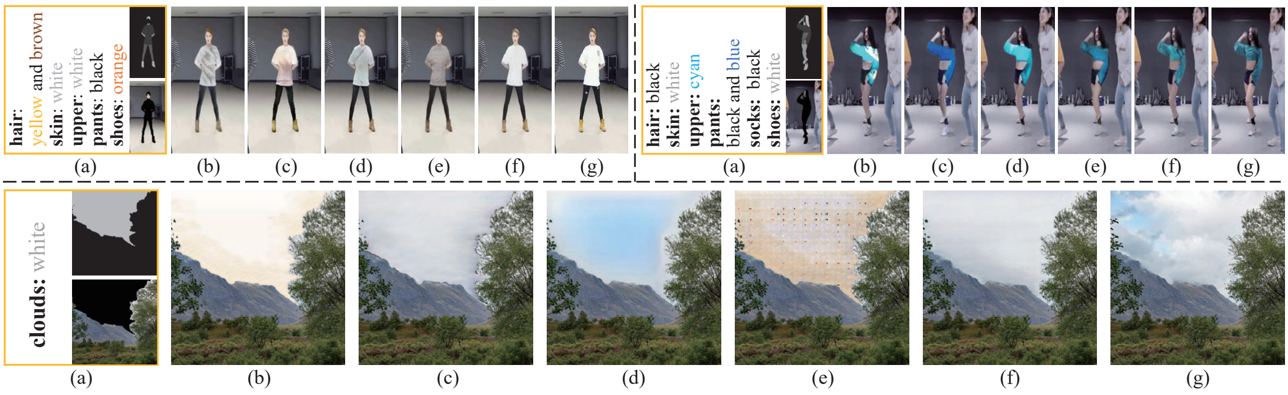


Figure 7. Ablation study with different variants of the proposed method on two datasets. (a) Input conditions: color (left), geometry (top right), and background (bottom right); texture is omitted here. (b) CMSM. (c) SINGLE. (d) W/o assist. (e) Two-phase. (f) Our results. (g) Original images. Note that we repaint all samples with the original conditions (the same as the original images). Please zoom in for details.

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.

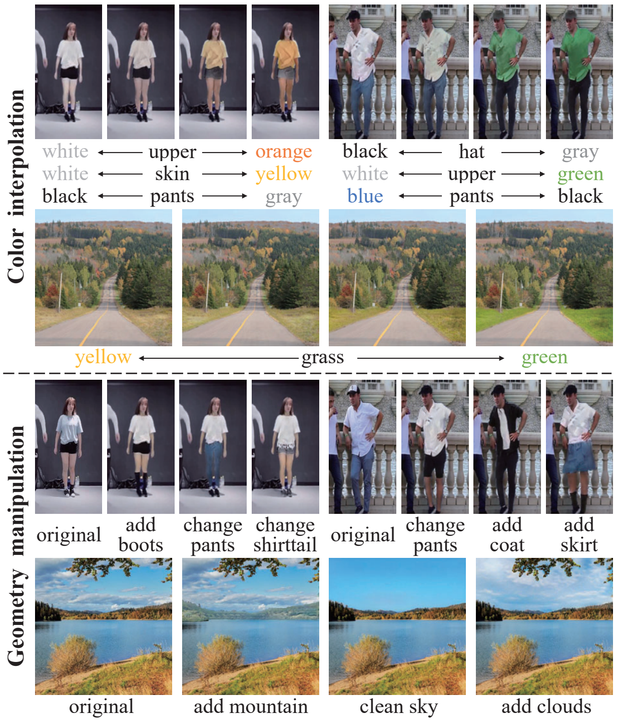


Figure 8. Controllability study results of UniCoRN with interpolated color and manipulated geometry. Please zoom in for details.

Acknowledgements

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

References

[1] Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden. Pyramid methods in image processing. *RCA engineer*, 29, 1984. 5

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 2

[3] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*. 1987. 5

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

[5] Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *CVPR*, 2019. 2, 5

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *TPAMI*, 40(4), 2017. 6

[7] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. DoveNet: Deep image harmonization via domain verification. In *CVPR*, 2020. 2

[8] Xuan Dong, Weixin Li, Xiaojie Wang, and Yunhong Wang. Learning a deep convolutional network for colorization in monochrome-color dual-lens system. In *AAAI*, 2019. 1

[9] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 5

[10] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[11] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code GAN prior. In *CVPR*, 2020. 2

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a nash equilibrium. In *NIPS*, 2017. 6

[13] Tianyu Hua, Hongdong Zheng, Yalong Bai, Wei Zhang, Xiao-Ping Zhang, and Tao Mei. Exploiting relationship for complex-scene image generation. In *AAAI*, 2021. 2

[14] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 1, 2

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[16] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *CVPR*, 2019. 3, 5

[17] Yuheng Li, Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Mixnmatch: Multifactor disentanglement and encoding for conditional image generation. In *CVPR*, 2020. 2

[18] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Self-supervised sketch-to-image synthesis. In *AAAI*, 2021. 2

[19] Varun Manjunatha, Mohit Iyyer, Jordan L Boyd-Graber, and Larry S Davis. Learning to color from language. In *NAACL*, 2018. 1

[20] Takeru Miyato and Masanori Koyama. CGANs with projection discriminator. In *ICLR*, 2018. 2

[21] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *ICML*, 2017. 2

[22] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 2, 3

[23] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 1

[24] Dario Pavllo, Aurelien Lucchi, and Thomas Hofmann. Controlling style and semantics in weakly-supervised image generation. In *ECCV*, 2020. 2, 6, 7

[25] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*. 2003. 1

[26] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *CVPR*, 2019. 2

[27] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 2

[28] Shashank Tripathi, Siddhartha Chandra, Amit Agrawal, Ambrish Tyagi, James M. Rehg, and Visesh Chari. Learning to generate synthetic data via compositing. In *CVPR*, 2019. 6

[29] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, 2017. 2

[30] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6

[31] 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. 2, 4, 5

[32] Shuchen Weng, Wenbo Li, Dawei Li, Hongxia Jin, and Boxin Shi. Conditional image repainting via semantic bridge and piecewise value function. In *ECCV*, 2020. 1, 2, 3, 6, 7, 8

[33] Shuchen Weng, Wenbo Li, Dawei Li, Hongxia Jin, and Boxin Shi. MISC: Multi-condition injection and spatially-adaptive compositing for conditional person image synthesis. In *CVPR*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[34] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. GP-GAN: Towards realistic high-resolution image blending. In *ACM MM*, 2019. 1, 2

[35] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *CVPR*, 2019. 1

[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

[38] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 1

[39] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 2

[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

[41] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. Deep image blending. In *WACV*, 2020. 1

[42] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 1

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

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

[45] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive temporal encoding network for video instance-level human parsing. In *ACM MM*, 2018. 6

[46] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. SEAN: Image synthesis with semantic region-adaptive normalization. In *CVPR*, 2020. 2, 6, 7