# Image Re-composition via Regional Content-Style Decoupling

Rong Zhang[1], Wei Li[2], Yiqun Zhang[1],
Hong Zhang[3], Jinhui Yu[1], Ruigang Yang[2], Weiwei Xu[1] *
[1]State Key Lab of CAD&CG, Zhejiang University. [2]Inceptio. [3]SenseTime Group Ltd.
cadzhangrong@zju.edu.cn,liweimcc@gmail.com,zyqlouise@zju.edu.cn,
fykalviny@gmail.com,jhyu@cad.zju.edu.cn,ryang@cs.uky.edu,xww@cad.zju.edu.cn,

## ABSTRACT

Typical image composition harmonizes regions from different images to a single plausible image. We extend the idea of image composition by introducing the content-style decomposition and combination to form the concept of image re-composition. In other words, our image re-composition could arbitrarily combine those contents and styles decomposed from different images to generate more diverse images in a unified framework. In the decomposition stage, we incorporate the whitening normalization to obtain a more thorough content-style decoupling, which substantially improves the re-composition results. Moreover, to handle the variation of structure and texture of different objects in an image, we design the network to support regional feature representation and achieve region-aware content-style decomposition. Regarding the composition stage, we propose a cycle consistency loss to constrain the network preserving the content and style information during the composition. Our method can produce diverse re-composition results, including content-content, content-style and style-style. Our experimental results demonstrate a large improvement over the current state-of-the-art methods.

## CCS CONCEPTS

• **Computing methodologies → Computational photography**.

## KEYWORDS

image composition, image manipulation, style transfer

**ACM Reference Format:**
Rong Zhang[1], Wei Li[2], Yiqun Zhang[1],, Hong Zhang[3], Jinhui Yu[1], Ruigang Yang[2], Weiwei Xu[1] . 2021. Image Re-composition via Regional Content-Style Decoupling. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475212

## 1 INTRODUCTION

Image composition is a long-lasting topic in image editing [4, 30, 33, 34]. A typical example is to crop a foreground region from a source image and paste it into the target image to generate a harmonic image. It is necessary to convert the local appearance of the cropped region to the pixel-level statics of the target so that the region can be compatible with the target. Traditional algorithms accomplish the conversion with color distribution matching [31]

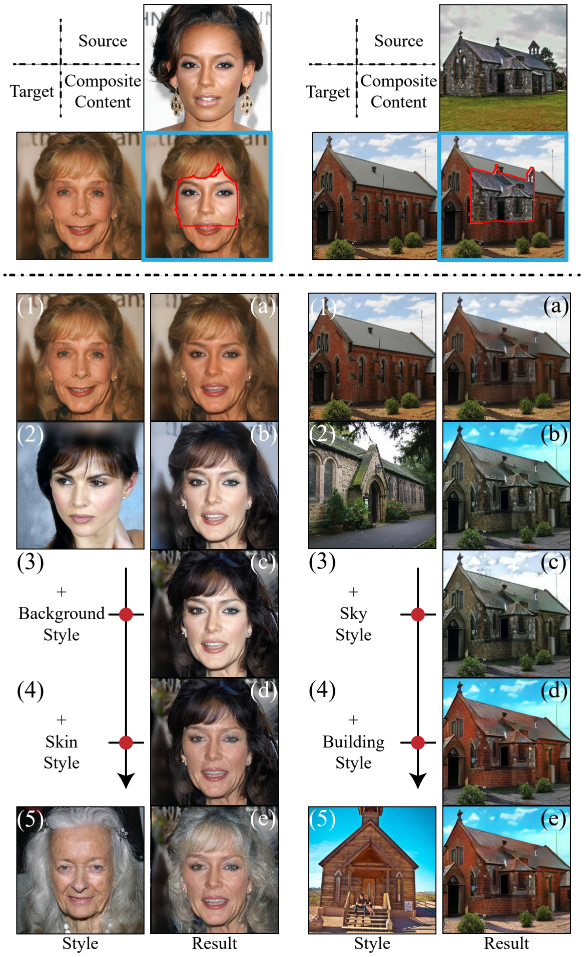


**Figure 1: Image re-composition of CelebAMask-HQ (left) and LSUN Church (right). Top two rows show content sources. The input contents are shown in blue boxes with red boundaries represent the composite masks. (a)-(e) are our re-composition results generated from input contents and corresponding styles (1)-(5) in previous columns. (a): content-content re-composition with composite contents and target styles (1); (b)/(e): content-style re-composition with composite contents and new styles (2)/(5); (c)/(d): style-style re-composition with composite contents and styles (3)/(4) composited from (2) and (5) region by region.**

*Corresponding author.

*MM '21, October 20–24, 2021, Virtual Event, China*
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10...$15.00
https://doi.org/10.1145/3474085.3475212

or gradient domain optimization [17, 30]. Recently, learning-based researches [24, 38, 40] utilize convolution neural networks (CNNs) to generate harmonized images directly. However, the scope of composition is still limited to composite regions from different images mostly.

In this paper, we further extend the concept of image composition to the harmonization of contents and styles decomposed from different images, which we call *image re-composition*. Our image re-composition is a unified solution to composite the content/style arbitrarily to generate realistic images. Here, content represents the overall spatial structure like edges, shapes, *etc.*, while style represents an image's local appearance such as color, texture, brightness, *etc.* Depending on the contents' or styles' sources, image re-composition can be divided into three categories (Fig. 1):

- content-content re-composition, e.g., classical image composition (or image-guided inpainting), which copies the content of a source image to a target image while preserving the target style;
- content-style re-composition, e.g., image style transfer, which composites the source image's style to the target content;
- style-style re-composition, e.g., regional style transfer, which combines part of source style with the target style.

Image re-composition has the chance to step into a more prominent stage of augmenting the limited amount of labeled images since it can generate more diversified images.

To achieve all the above categories of applications in one unified framework, we propose an end-to-end image re-composition network from the perspective of content-style decomposition. Given an image, the network decouples contents from highly twining styles so that the content and style could be composited arbitrarily as individual components. A primary challenge is how to decompose an image to content and style. The most recent work swapping AutoEncoder (AE) [28] produces very compelling results by decomposing two images to contents and styles implicitly and generating hybrid images after swapping their styles. Nevertheless, swapping AE suffers from artifacts when content and style are highly coupled. Instead of implicitly decomposing content and style, we revisit the definition of style and design an enhanced region-based content decompositor with whitening normalization to explicitly distill clean content from texture-rich images. On the other hand, almost every real-world image has different types of elements or textures in different regions. Mixing them would cause undesired effects in the re-composition result. We use regional style representation to keep the uniqueness of different elements.

After the decomposition, how to bond styles back with their corresponding locations is another critical problem. An alternative choice is using Semantic Region-Adaptive Normalization (SEAN) block [42] to restore the regional styles to contents in semantic regions. But since the input content of [42] is semantic segmentation labels, it is not able to precisely control structures of different images and composite them to a new image like image composition. Besides, the SEAN block extracts the modulation parameters from both the style matrix and the segmentation mask, making it rely on the fine-scale segmentation map. In our approach, the contents supply detailed structure information, and our regional re-composition content-style composition layer extracts modulation parameters only based on the styles so that our network is compatible with coarse labels.

In summary, we propose a well-designed architecture that can blend content and style arbitrarily to support seamless regional image composition. Our contribution is as follows:

- We introduce an end-to-end framework to re-composite two or more images to generate new images with desired styles while preserving corresponding content information.
- We introduce a region-based content decompositor with whitening normalization to further decompose the content& style and redesign the training scheme to cooperate with it.
- Our network is able to achieve three kinds of re-composition of content and style in a unified framework. We design some metrics to evaluate the results. The experiments demonstrate a significant improvement over state-of-the-art methods.

## 2 RELATED WORK

*Image-to-Image Translation.* is the task of translating an image into another with different styles. Early work utilized convolution neural networks (CNN) to map images between two predefined domains [16, 41]. However, these methods are hard to synthesize images with a large number of styles in the dataset. Recently, with the help of feature modulation technique called Adaptive Instance Normalization (AdaIN) [14], the style based models are able to generate high quality photo-realistic images regardless of the resolution. Nevertheless, this family of methods is limited to the randomly generated images, unable to tackle with specific real images. To address this issue, Abdal *et al.* [1, 2] proposed to map the real images back into the latent space, and then edit the images by manipulating the latent codes. However, this approach suffers from both a slow optimization process and low reconstruction quality. Recent work [7, 8, 15] has shown that the performance of image translation can be improved by decomposing the images into content and style. Park *et al.* [28] designed an autoencoder network and introduced a co-occurrence patch discriminator to enforce the disentanglement of the two independent components.

Almost all the above methods focus on the full image-to-image translation. The other line of research explores methods to enable semantic image synthesis. SPADE [27] adopted the segmentation mask to manipulate the labeled regions. However, the full image is edited by one style, which is insufficient for precise control. Later, Zhu *et al.* [42] introduced the SEAN module to improve the per-region style encoding, which harvests the style of semantic regions. Chen *et al.* [5] proposed Semantic Instance Wised StyleGAN to translate free-viewpoint semantic maps to images.

*Image Composition.* is also called image harmonization, aiming at generating a synthesized image hybrid by pasting the foreground of the guided image into the background of the target image. Therefore, the main challenge is to make the foreground part compatible with the background, especially for the boundary region. Traditional image composition models focus on match the pixel-level statistics of the two images, such as image blending [12], matching the color distribution [31], utilizing gradient-domain composition [17], and matching images into the carefully crafted color templates [9]. Considering both the low-level statistics and the image content, many works around visual realism are proposed [22, 34]. Recently,

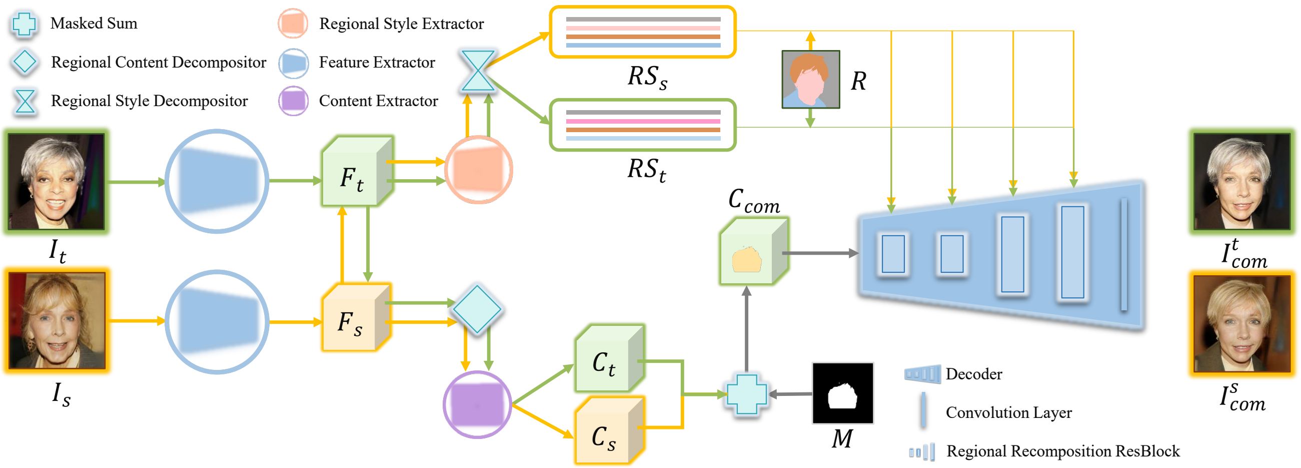


**Figure 2: Our network structure for image re-composition. Given a source image $I_s$ and a target image $I_t$, The feature extractor encodes them into $F_s$ and $F_t$. Then we use a content extractor and a regional content decompositor to distill clean content codes $C_s, C_t$ without style information. The content codes are masked summed to the composited content $C_{\text{com}}$. In the meantime, the regional style extractor and style decompositor use a composited region map R extract the regional style codes $RS_s$ and $RS_t$ for both images. The decoder with region re-composition block can recompose $C_{\text{com}}$ and any of the style code to generate composited images $I^t_{\text{com}}$ with target image's style and $I^s_{\text{com}}$ with source image's style.**

several works [3, 33, 40] proposed to use CNN to directly produce the composited images, boosting the image quality to a new level. Other work further explored novel network architecture techniques like attention module [10] to improve the performance.

*Style Transfer.* Our work is also related to style transfer explored in [6, 26, 36], in which the style information is strictly injected into the content of the given image. [32] achieves regional style transfer between different classes using spatial conditional batch normalization conditioned on predefined classes. But the reconstructed content is not accurate. Besides stressing the content maintaining and the style formulation, our model could incorporate the high-level semantic features into the given image. With the whitening technique in style transfer, our model improves image composition by encouraging the disentanglement of content and style.

## 3 FORMULA DEFINITION OF IMAGE RE-COMPOSITION

In image composition, the images' styles may vary in color, illumination, brightness, *etc.* Compositing the images' contents together and seamlessly blending their styles is quite challenging. We use image re-composition to reduce the influence of the style. It has three steps: decomposition, manipulation, and re-composition. The decomposition module $E$ extracts a feature map $F$ from a given image $I$ and encodes it to a content code $C$ and a style code $S$, which can be represented as $C, S = E(I)$. The re-composition is the reverse process of decomposition. The re-composition decoder/generator reconstructs a realistic image $I'$ by composing the content code and style code $I' = G(C, S)$.

The manipulation involves the composition of two or more images' contents or styles. Considering a two-image situation with a source image $I_s$ and a target image $I_t$, we can combine the source content $C_s$ with the target content $C_t$ or the source style $S_s$ with the target style $S_t$. $C_{\text{com}} = C_t \odot M + C_s \odot (1 - M)$ is the manipulated content code under the control of mask $M$. $\odot$ represents element-wise multiplication. Similarly, a composited style code $S_{\text{com}}$ can be generated from $S_s$ and $S_t$.

Regrading the re-composition module, $G$ should be able to generate realistic images from the manipulated content $C_{\text{com}}$ or style $S_{\text{com}}$. For example:

$$\begin{aligned} I^t_{\text{com}} &= G(C_{\text{com}}, S_t), && \text{content-content re-composition} \\ I^s_{\text{com}} &= G(C_{\text{com}}, S_s), && \text{+content-style re-composition} \\ I^{\text{com}}_{\text{com}} &= G(C_{\text{com}}, S_{\text{com}}), && \text{+style-style re-composition} \end{aligned} \tag{1}$$

The fake image $I^t_{\text{com}}$ represents a fake hybrid image with the same style of $I_t$ and composited content of $I_s$ and $I_t$. The subscript of $I^t_{\text{com}}$ shows where the content comes from, while the superscript shows the style's source.

## 4 APPROACH

Based on the above section's objectives, the recent deep image manipulation algorithm swapping AE is the most related research. In the following, we revisit the concept of content-style separation/aggregation in Swapping AE and describe the main components in our network, including how to disentangle the style information from the image contents Sec. 4.1, how to recompose them together Sec. 4.2, and how to train the network Sec. 4.3.

Swapping AE decomposes an image to content and style with a feature swapping training strategy. Then they use a co-occurrence patch discriminator to enforce $I^s_t$ and $I_s$ with the same low-level patch distribution to constrain them to have the same style. This is an implicit decomposition in terms of content and style. Based on features from such decomposition, we suppose to get a realistic composited image $I^s_{\text{com}}$. However, in our experiments, swapping AE may yield artifacts on the image composition task (Fig. 7). The reasons can be in two-folds: 1). The content code is still mixed with

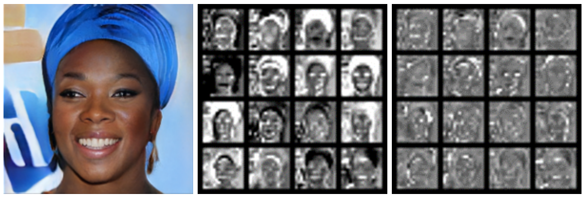


**Figure 3: Visualization of the whitening operation. From left to right: the input image, features before and after the whitening.**

style information so that $I^{s}_{\text{com}}$ preserves the partial style of $I_s$. 2). $G$ mapped the same object with a different local style.

To fix these issues, we decompose the content & style in the region-level with a regional content decompositor (RCD) and a regional style decompositor (RSD). Then the content and style are recomposed together by regional re-composition layers. We also redesign the training scheme to cooperate the new modules.

## 4.1 Regional Content Decompositor with Whitening Normalization

To further disentangle the content code, we intend to transform the $C$ to $C'$ so that $C'$ only contains the content information. In the meanwhile, other information $\Delta(C, C')$ is transferred to $S$:

$$C', S + \Delta(C, C') = E(I) \tag{2}$$

There are two ways to achieve this objective: finding content $C'$ directly or removing style information $\Delta(C, C')$ from $C$. How to define the content and style? In [35], Yang et al. used the high-frequency features such as edges as the content. However, the "high frequency" is hard to define to get $C'$ directly. In the style transfer area, the researchers [11, 18] use the correlation between features (covariance matrix) to represent the style information. A feature map with normalized covariance has similar styles in different spatial locations. Based on that, we can eliminate styles from contents via constraining the content features' covariance matrix to be identity, a.k.a. a whitening normalization operation. The content decomposition operation can be defined as follows:

$$\begin{aligned} \mathbf{U}, \mathbf{\Lambda}, \mathbf{V} &= SVD(ff^{\mathsf{T}}) \\ f' &= \mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^{\mathsf{T}} f, \end{aligned} \tag{3}$$

where $f$ is the centralized extracted feature map, $ff^{\mathsf{T}}$ is the covariance matrix. $SVD$ is a singular value decomposition operator. $\mathbf{U}$ and $\mathbf{\Lambda}$ are the orthogonal eigenvectors matrix and diagonal eigenvalues matrix of the covariance matrix respectively. $f'$ is the feature map after whitening which satisfying $f' f'^{\mathsf{T}} = I$, $I$ is an identity matrix. Furthermore, to normalize the covariance in individual regions, the RCD can be defined as:

$$\begin{aligned} \mathbf{U}^n, \mathbf{\Lambda}^n, \mathbf{V} &= SVD(f^n f^{n\mathsf{T}}) \\ f' &= \sum_{n=0}^{N-1} (\mathbf{U}^n \mathbf{\Lambda}^{n^{-\frac{1}{2}}} \mathbf{U}^{n\mathsf{T}} f^n) \odot R^n, \end{aligned} \tag{4}$$

where $R$ is a segmentation mask with $N$ classes, $f^n$ is the feature map of the $n$th class. After the whitening normalization operation, high-frequency edges are preserved to represent the contents. Fig. 3 shows the visualized feature maps before/after whitening. The mean/covariance statistics of the features are removed from the content branch and can be re-learned in the style branch in training.

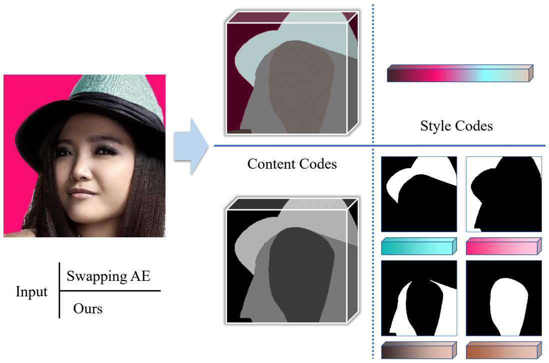


**Figure 4: The illustration shows the difference between the content codes and style codes decomposed by the swapping AE and our method. The content code of swapping AE (the top row of the middle column) still contains some style information (the color is not removed thoroughly), while our method can get clean content codes. Besides, the style code of the swapping AE (the top row of the right column) is entangled together to encode the whole image's style. On the contrary, our style codes are decomposed into regional ones. Each of them encodes a class of styles in the same region.**

## 4.2 Regional Content-Style Composition

Swapping AE [28] extracts a global style code by a global average pooling and injects the style to the content using the weight modulation layer from StyleGANv2 [20]. Due to the global style encodes all information in the whole image, this modulation is not semantic-aware. However, almost every real-world image is with different styles in different regions. Mixing them would cause undesired effects in the composition results (Fig. 5).

To deal with this issue, the RSD locates the regional styles $RS$ with a semantic segmentation map by regional average pooling. Then each image can be reconstructed with the regional content code $C'$ and $RS$.

$$\begin{aligned} RS &= RSD(F_s, R) \\ I' &= G(C', RS, R) \end{aligned} \tag{5}$$

where $F_s$ is the feature map in the style branch.

With the regional content code and style code, we need to recompose them to a new image. In [42], the authors use region-based mean and variance codes and instance normalization to achieve

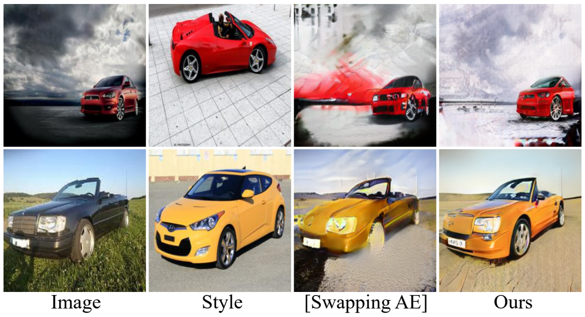


**Figure 5: Swapping AE failure cases.**

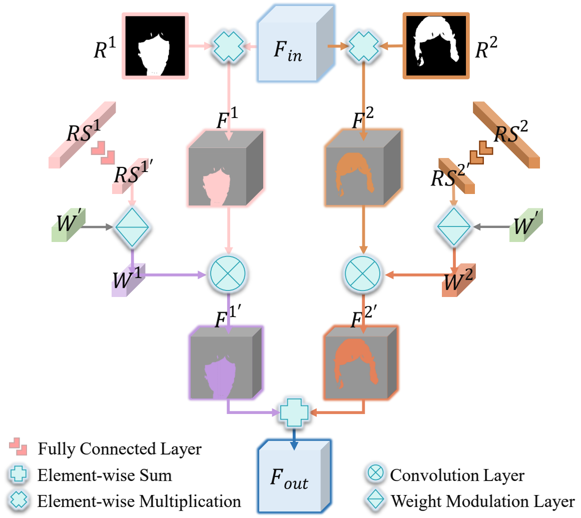


**Figure 6: Details of the regional re-composition layer. It extracts parameters $RS'$ from regional style code $RS$ to modulate original weights $W'$ to regional weights $W$. Then all convoluted feature maps are merged region by region.**

region editing. Since the SEAN block extracts the modulation parameters from both the style matrix and the segmentation mask, the composited image is bonded too tightly with the segmentation map to hallucinate realistic boundary areas along with the composted masks (col #6 of Fig. 7). Different from them, we directly extend the weight modulation layer to a region-based modulation layer to avoid this issue. With input features $F_{in}$, original weights $W'$, the regional re-composition layer is defined as the following equations:

$$w^n_{ijk} = \frac{rs_i^{n'} \cdot w'_{ijk}}{\sqrt{\sum_{i,k} {w'_{ijk}}^2 + \epsilon}}, w'_{ijk} \in W' \tag{6}$$

$$F_{out} = \sum_{n=0}^{N-1} (F_{in} \odot R^n) \otimes W^n \odot R^n \tag{7}$$

where $i, j, k$ are the indexes corresponding to the input channel, output channel, and spatial dimension respectively. $w \in W$ is the weight after modulation. $rs_i^{n'}$ is the $i$th extracted modulation parameter from the $n$th style code. $F_{out}$ is the output features. $\otimes$ and $\odot$ are convolution operator and element-wise multiplication. Fig. 6 shows the details of the regional re-composition layer.

### 4.3 Loss Function and Training Scheme

Our network consists of four parts: a feature extractor $E_F$, a content code extractor $E_C$, a regional style code extractor $E_S$, and a decoder/generator $G$. Fig. 2 shows the overall structure of the proposed network.

Given the input images $I_s, I_t$, the content extractor $E_C$ and $E_S$ extract their content code $C_s, C_t$ and regional style code $RS_s, RS_t$ respectively. Then the decoder generates a reconstructed image $I_t^t = G(C_t, RS_t, R_t)$ and a fake hybrid (style-swapped) image $I_t^s = G(C_t, RS_s, R_t)$. The training loss of the generator $L_g$ consists of three parts: a reconstruction loss $L_{\text{rec}}$, an adversarial loss $L_{\text{adv}}$ and a cycle consistency loss $L_{\text{cycle}}$.

$$\begin{aligned} L_g &= L_{\text{rec}} + \alpha L_{\text{adv}} + \beta L_{\text{cycle}} \\ L_{\text{rec}} &= ||I_t - G(C_t, RS_t, R_t)|| \\ L_{\text{adv}} &= L_{\text{GAN, rec}} + L_{\text{GAN, swap}} + L_{\text{CooccurGAN}} \end{aligned} \tag{8}$$

where $\alpha$ and $\beta$ are the weights of the losses. A global discriminator $D$ and a co-occurrence discriminator $D_{patch}$ are used to training the network. We adopt the same adversarial loss and discriminator loss following [28].

*Cycle consistency losses.* Our cycle consistency loss aims to enforce the generated fake hybrid image to preserve the target image's exact content and the source image's same regional styles. $L_{\text{cycle}}$ is defined as:

$$L_{\text{cycle}} = L_{\text{cycle, content}} + L_{\text{cycle, styles}} \tag{9}$$

$L_{\text{cycle, content}}$ is the $L1$ loss between $C_t$ and the content of $I_t^s$.

$$L_{\text{cycle, content}} = |C_t - E_C(E_F(I_t^s))| + |C_t - E_C(E_F(I_t^t))| \tag{10}$$

For $L_{\text{cycle, styles}}$, we use the cosine distance to measure the similarity of $RS_s$ and $I_t^s$'s regional styles $RS_{hyb} = E_S(E_F(I_t^s, R_t)$.

$$L_{\text{cycle, styles}} = \mathbb{E}[1 - \frac{RS_s^n \cdot RS^n_{\text{hyb}}}{||RS_s^n||_2 \cdot ||RS^n_{hyb}||_2}] \tag{11}$$

$\mathbb{E}$ is a mean operator.

*Training Scheme.* In style transfer, the whitening operation is often used in the inference process and does not influence the network training. In our method, with the embedded whitening operation, the information related to the style in the content branch will be abandoned. We need to force the style branch to pick up the missing information to reconstruct the image. We design a two-stage training scheme to incorporate the regional modules and the whitening normalization operator.

In the first stage, we train our model with the regional modules from scratch to enhance the regional style representation. The network is trained with the swapping strategy in [28]. This stage helps to locate a proper initialization for the whitening normalization.

In the second stage, the RCD is plugged in. We fix the feature extractor and only optimize the style extractor, content extractor and other modules to force the network to transform the residual style information in the content to the style code.

## 5 EXPERIMENTS

### 5.1 Implementation Details

We implement the network with PyTorch [29] and train it on two NVIDIA V100 GPUs. The weights of adversarial loss and cycle consistency loss are all 1.0. The learning rates for the generator and discriminators are 0.002. We use the ADAM optimizer with $\beta_1 = 0, \beta_2 = 0.99$. The experiments are conducted on CelebAMask-HQ [19, 23, 25], LSUN Church[37] and Cars[21]. More details can be referred to in the supplementary materials.

### 5.2 Comparisons

*Metrics.* The proposed approach aims to re-compose input images' contents and styles to generate a new image close to the desired style. Obviously, there is no ground truth for the generated

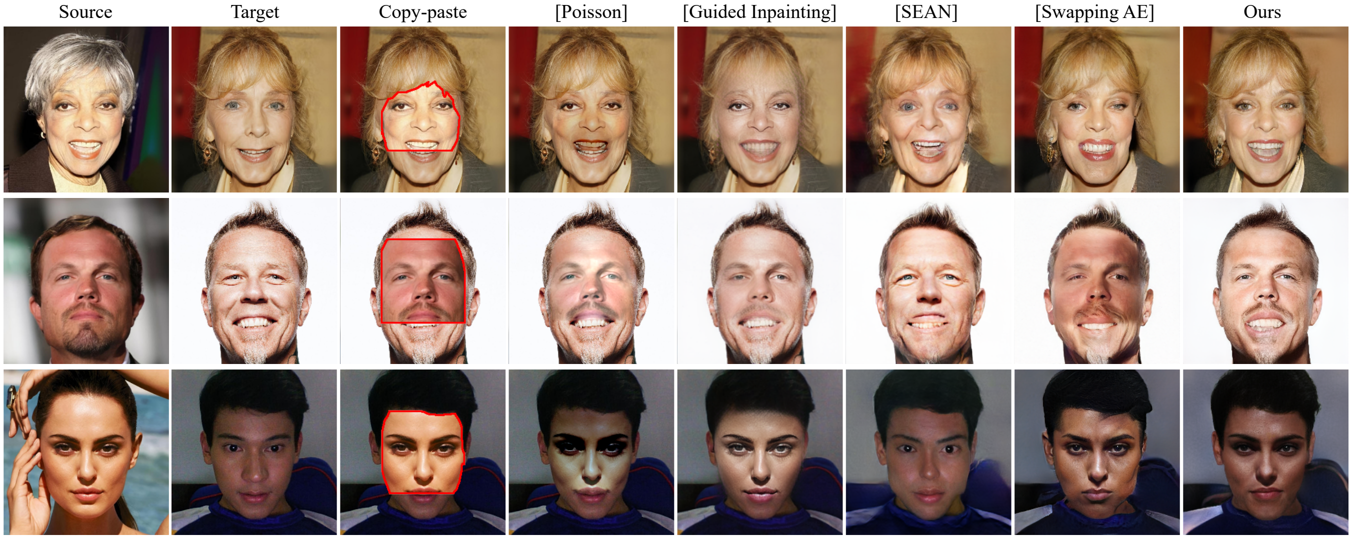


**Figure 7: Comparison with baselines: CelebAMask-HQ dataset. Copy-paste indicates that RGB values of the source in the masked region are directly copied into the target. Poisson has artifacts along the boundary. GIP fails to reconstruct the details. SEAN is not able to preserve the center content of the source image. The style inconsistency occurs in Swapping AE.**

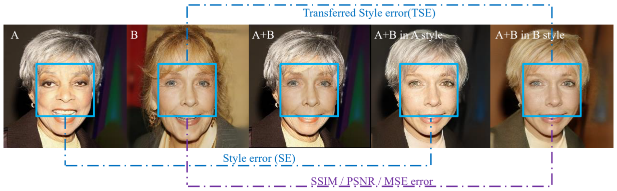


**Figure 8: Evaluation metrics.**

image. So we introduce a set of metrics for evaluation. Given two images $I_s, I_t$, we combine their contents with masked-sum and generate the composition results in source style $I^s_{\text{com}}$ and target style $I^t_{\text{com}}$ respectively. $I^t_{\text{com}}$ and $I_t$ have the same style but different content in the mask. We can calculate the style error (SE) [18] between them to measure whether the desired style is well-preserved. $I^s_{\text{com}}$ and $I_s$ have the same style and content in the mask region. We can use common supervised metrics such as peak signal-to-noise ratio (PSNR), structural similarity (SSIM) and mean squared error (MSE) as well as transferred style error (TSE) to evaluate their distance. We also use the Fréchet Inception Distance (FID) [13] to measure the image quality and diversity. The metrics are shown in Fig. 8.

*Baselines.*

- Poisson Blending (Poisson) [30] is a classic image composition algorithm with gradient domain fusion.
- SEAN [42] uses GANs to generate synthetic images by "adding" realistic styles to semantic masks. To do image re-composition, we first combine the semantic masks together and transfer the target image style to the mask.
- Guided Inpainting (GIP) [39] achieves image inpainting by pulling content from one image to another and regenerating boundary regions. To calculate the evaluation metrics, we apply the SOTA photo-realistic style transfer algorithm WCT2 [36] to generate $I^s_{com}$.
- Swapping AE [28] implicitly maps the image feature to structure space and texture space. It can generate image re-composition results similar to our method except for regional manipulations.

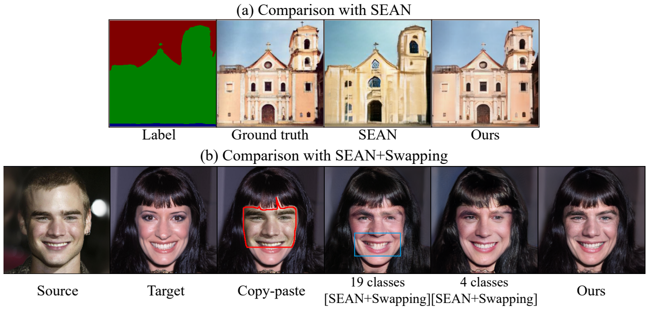

(a) Comparison with SEAN
Label
Ground truth
SEAN
Ours
(b) Comparison with SEAN+Swapping
Source
Target
Copy-paste
19 classes [SEAN+Swapping]
4 classes [SEAN+Swapping]
Ours

**Figure 9: Comparison with SEAN and SEAN+Swapping.**

*Results and Analysis.* Fig. 7 and Fig. 11 show visual comparisons on image re-composition between the baseline methods and our approach. Poisson Blending may fail to preserve the morphological appearance of the image, especially on the boundary, as it only considers the smoothness of the gradient domain. It is not capable of fixing the inconsistency along the boundary. In contrast, GAN models have the generative ability to correct the mismatched boundaries to some extent.

GIP can locate the boundary area which is unmatched with the surrounding context and generate new content in this area to synthesize different images. But it fails to preserve details and keeping the image style consistent.

Swapping AE can composite the contents of input images as it extracts content and style code separately. Although injecting the target style to the composited content code, the source image's style still can be found in the second to the last column of Fig. 7. It proves that the content code in swapping AE mixes with style information.

As an image-to-image translation algorithm, SEAN is good at maintaining the image style. However, the contents from different

| Method | SSIM | PSNR | MSE | TSE | SE | FID |
|---|---|---|---|---|---|---|
| Poisson | - | - | - | - | 1.174 | 10.27 |
| SEAN | **0.654** | 20.97 | 0.0093 | 0.489 | 0.723 | 17.43 |
| GIP-WCT2 | 0.624 | 14.827 | 0.042 | 1.1108 | 1.1881 | 19.21 |
| Swapping AE | 0.590 | 20.908 | 0.0094 | 0.497 | 1.0274 | 11.63 |
| SEAN+Swapping | 0.590 | 20.067 | 0.0114 | 0.5871 | 0.81 | 14.27 |
| Ours | 0.6431 | **21.773** | **0.0075** | **0.3639** | **0.7105** | **10.16** |

**Table 1: Quantitative comparison with different methods. For SSIM and PSNR, higher is better. For MSE, SE, TSE and FID, lower is better. SE and TSE are in the range of $10^{-4}$.**

| Data | Our | Swapping AE | Poisson | GIP | SEAN |
|---|---|---|---|---|---|
| CelebA | **34.5%** | 12.1% | 11.4% | 17.8% | 24.2% |
| Church | **61.4%** | 17.2% | 12.7% | 8.7% | - |

**Table 2: Human perceptual study result.**

images can not be preserved simultaneously since it uses semantic labels as contents. In the case of Fig. 7 (col #6), a good composited image should make users identify the face as the person in the source image while SEAN results are always with the target image's characteristics. On the other hand, the SEAN block heavily relies on fine-scale semantic segmentation as it extracts modulation parameters from two branches: the style matrix and the segmentation mask. In the Church dataset, the main buildings have many details and no fine-scale semantic labels, making SEAN fail to reconstruct the building Fig. 9 (a). Therefore, we do not present SEAN results of the Church dataset in Fig. 11.

Considering the SEAN module is also a regional re-composition block, we conduct an experiment to compare it with our block. We replace our regional re-composition layer with SEAN block and train the network with the swapping strategy to composite the contents. We represent it as SEAN+Swapping. However, when we use semantic labels of 19 classes, the outputs strictly follow the segmentation masks, making the composted boundary areas not realistic. When we use four-category labels (same as ours), the reconstruction is not accurate for a smaller nose and missing eyebrow Fig. 9 (b). Our regional re-composition is superior to SEAN+Swapping for fewer artifacts, better details and better boundary hallucination.

In summary, comparing to the baseline methods, our method achieves better results in three aspects:

- Content preserving. Our method can composite the source image's content to the target image by manipulating the separated content code.
- Style consistency. We propose the whitening normalization, regional representation and cycle consistency loss to ensure the whole generated image keeps a similar style to the target image.
- Natural boundary. As a generative model, our network can process misaligned regions and hallucinate realistic boundaries.

Tab. 1 can verify that our method outperforms others on most metrics.

We also conduct a human perceptual study to evaluate the realism of results further. We randomly select 40 images from CelebAMask-HQ and Church. The source, target, and composite images of different methods are presented simultaneously. 63 people from Amazon Mechanical Turkers and Tencent are recruited to choose the most realistic result. More than 2400 samples are collected. Tab. 2 shows our method is superior to others, especially on the Church dataset of high diversity.

| Method | SSIM | PSNR | MSE | TSE | SE |
|---|---|---|---|---|---|
| Swapping AE | 0.590 | 20.908 | 0.0094 | 0.497 | 1.0274 |
| +cycle loss | 0.603 | 21.142 | 0.0090 | 0.4843 | 0.9942 |
| +whitening | 0.618 | 21.339 | 0.0082 | 0.4639 | 0.815 |
| +regional style | 0.637 | 21.767 | 0.0077 | **0.3529** | 0.7986 |
| Ours | **0.6431** | **21.773** | **0.0075** | 0.3639 | **0.7105** |

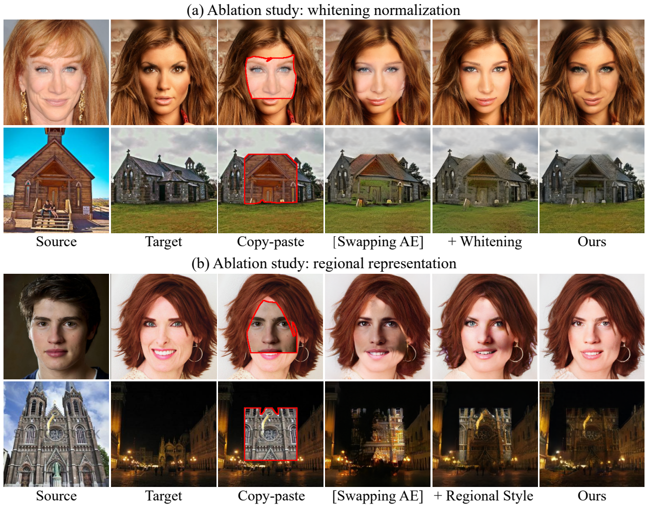
**Table 3: Ablation study metrics. For SSIM and PSNR, higher is better. For MSE, SE, and TSE, lower is better. SE and TSE are in the range of $10^{-4}$.**

## 5.3 Ablation Study

To explore the effects of our algorithm's different parts, we conduct several ablation experiments about cycle consistency loss, the whitening normalization operator, and regional style swapping. Tab. 3 shows the evaluation metrics of the ablation study.

All our modules can improve most metrics. Adding the cycle consistency loss can improve the SSIM and reduce the TSE as it can constrain the network to preserve the content and style during the style transfer process. Adding whitening normalization in the global region can improve SSIM from 0.603 to0.618 and reduce SE from 0.994 to 0.815. With the help of feature whitening, the content codes contain less style-related information. The content composition will cause minor style inconsistency, especially when the input images' styles differ significantly. Fig. 10 (a) shows some composition results with/without whitening operation. It can be seen that adding whitening can process more challenging cases with very different input images and generate results with fewer artifacts.

Besides, adding the regional style swapping can also gain further promotion of the metrics. It can utilize the semantic information to locate different objects to extract regional styles and guide the style code injecting process explicitly to ensure similar contents have


(a) Ablation study: whitening normalization

(b) Ablation study: regional representation

**Figure 10: Ablation study. "+Whitening": results with whitening only. "+Regional Style": results with regional representation only. "Ours": results with all modules.**

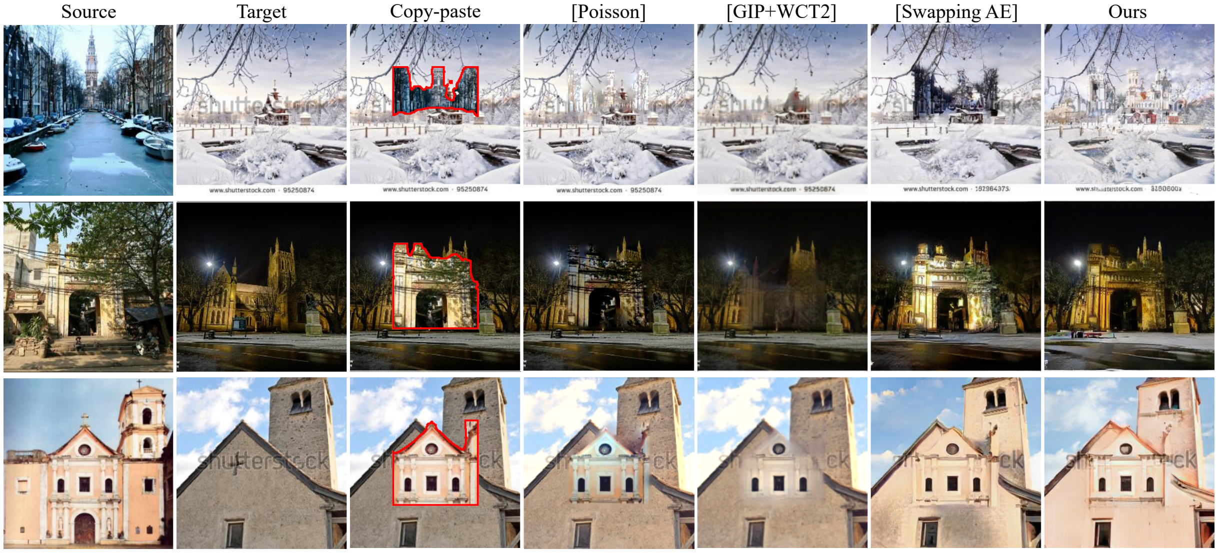


**Figure 11: Comparison with baselines: Church dataset. The last row shows the re-composited result in the source style.**

similar styles. In this way, even if the content code has changed after the content composition, the network still knows where to inject the style code. Fig. 10 (b) shows the comparison results of baseline, adding regional style swapping, and our entire algorithm. In this figure, due to the input images having quite diverse skins, Swapping AE's results suffer severe inconsistency. Introducing regional style swapping reduces the artifacts by transferring the same regional style code to the same content region.

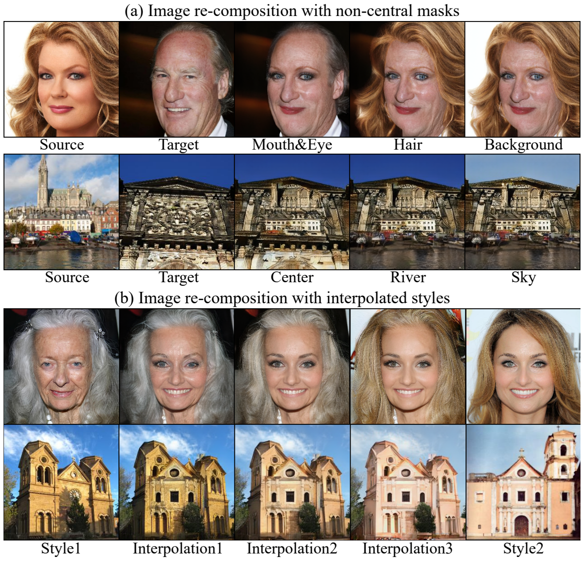
## 5.4 Additional Results

As we state in Sec. 1, we decompose every image into region-based contents and styles so that we can randomly or artificially combine them to achieve more diverse re-composition images, such as content + content/ content + style or style + style. We will show additional image re-composition results in the following.

*Results with non-central masks.* Our network can achieve content-content re-composition while preserving the image style, which is similar to the classic image composition. Furthermore, the styles also can be composited simultaneously. Fig. 12 (a) shows the composition results with non-central masks. We composite the source image's content and style to the target image region by region.

*Results with style interpolation.* Regarding the content-style re-composition, our network can composite any style code to the content code, just like the traditional style transfer. The style code also can be linearly interpolated with the source and target style. Fig. 12 (b) shows some results with interpolated styles.

*Failure cases.* Our method may be influenced when the objects of the input images are in significantly different poses or when a semantic class in the source does not exist in the target. Details can be referred to in the Supplementary Materials.


(a) Image re-composition with non-central masks

(b) Image re-composition with interpolated styles

**Figure 12: Additional re-composition results. (a). Compose regional content and style to the target region by region. (b). Recompose a composite content with interpolated styles.**

## 6 CONCLUSION

We have developed an enhanced content-style decomposition method for regional image re-composition. It incorporates the regional whitening operation to obtain a more thorough content-style decomposition which substantially improves the re-composition results. To handle the variation of structure and texture of different objects in an image, we design our pipeline to support regional feature swapping and achieve region-aware content-style decomposition. Experimental results verify that our method can produce diverse seamless re-composition results in a unified framework, including content-content, content-style and style-style.

## ACKNOWLEDGMENTS

We would like to thank anonymous reviewers for their constructive comments. Weiwei Xu is supported by NSFC (No. 61732016). Jinhui Yu is supported by NSFC (No. 61772463).

# REFERENCES

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?. In *ICCV*.

[2] R. Abdal, Y. Qin, and P. Wonka. 2020. Image2StyleGAN++: How to Edit the Embedded Images?. In *CVPR*.

[3] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. 2020. Compositional gan: Learning image-conditional binary composition. *International Journal of Computer Vision* 128, 10 (2020), 2570–2585.

[4] Peter J Burt and Edward H Adelson. 1983. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics (TOG)* 2, 4 (1983), 217–236.

[5] Anpei Chen, Ruiyang Liu, Ling Xie, and Jingyi Yu. 2020. A free viewpoint portrait generator with dynamic styling. *arXiv preprint arXiv:2007.03780* (2020).

[6] Tai-Yin Chiu. 2019. Understanding Generalized Whitening and Coloring Transform for Universal Style Transfer. In *ICCV*.

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *CVPR*.

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*.

[9] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. 2006. Color harmonization. In *SIGGRAPH*.

[10] Xiaodong Cun and Chi-Man Pun. 2020. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing* 29 (2020), 4759–4771.

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

[12] Nuno Gracias, Mohammad Mahoor, Shahriar Negahdaripour, and Arthur Gleason. 2009. Fast image blending using watersheds and graph cuts. *Image and Vision Computing* 27, 5 (2009), 597–607.

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

[14] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.

[15] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal Unsupervised Image-to-image Translation. In *ECCV*.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[17] Jiaya Jia, Jian Sun, Chi keung Tang, and Heung yeung Shum. 2006. Drag-and-drop pasting. In *SIGGRAPH*.

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8110–8119.

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*. 554–561.

[22] Jean-Francois Lalonde and Alexei A Efros. 2007. Using color compatibility for assessing image realism. In *ICCV*.

[23] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5549–5558.

[24] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. 2018. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9455–9464.

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*. 3730–3738.

[26] Ming Lu, Hao Zhao, Anbang Yao, Yurong Chen, Feng Xu, and Li Zhang. 2019. A closed-form solution to universal style transfer. In *ICCV*.

[27] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2337–2346.

[28] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. 2020. Swapping Autoencoder for Deep Image Manipulation. *Advances in Neural Information Processing Systems* 33 (2020).

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[30] Patrick Pérez, Michel Gangnet, and Andrew Blake. 2003. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*. 313–318.

[31] F. Pitie, A.C. Kokaram, and R. Dahyot. 2020. N-dimensional probability density function transfer and its application to color transfer. In *ICCV*.

[32] Ryohei Suzuki, Masanori Koyama, Takeru Miyato, Taizan Yonetsuji, and Huachun Zhu. 2018. Spatially controllable image synthesis with internal representation collaging. *arXiv preprint arXiv:1811.10153* (2018).

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

[34] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. 2012. Understanding and improving the realism of image composites. *ACM Transactions on graphics (TOG)* 31, 4 (2012), 1–10.

[35] Yanchao Yang and Stefano Soatto. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4085–4095.

[36] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. 2019. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE International Conference on Computer Vision*. 9036–9045.

[37] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365* (2015).

[38] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. 2019. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3653–3662.

[39] Yinan Zhao, Brian Price, Scott Cohen, and Danna Gurari. 2019. Guided image inpainting: Replacing an image region by pulling content from another image. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1514–1523.

[40] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A Efros. 2015. Learning a discriminative model for the perception of realism in composite images. In *ICCV*.

[41] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *ICCV*.

[42] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. SEAN: Image Synthesis with Semantic Region-Adaptive Normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5104–5113.