# SCS-Co: Self-Consistent Style Contrastive Learning for Image Harmonization

Yucheng Hang[1,*], Bin Xia[1,*], Wenming Yang[1,2,†], Qingmin Liao[1,2]
[1] Shenzhen International Graduate School, Tsinghua University, China
[2] Department of Electronic Engineering, Tsinghua University, China
{hangyc20, xiab20}@mails.tsinghua.edu.cn, {yang.wenming, liaoqm}@sz.tsinghua.edu.cn

(a) Composite (b) DoveNet [5] (c) iS$^2$AM [35] (d) RainNet [24] (e) Ours (f) Real (Ground Truth)

Figure 1. Comparison with other methods. Thanks to the proposed SCS-Co, our method can obtain more explicit distortion knowledge from dynamically generated negative samples, and further jointly constrain the solution space from two aspects of the foreground self-style and foreground-background style consistency. Coupled with BAIN, our method produces a more photorealistic harmonized result.

## Abstract

*Image harmonization aims to achieve visual consistency in composite images by adapting a foreground to make it compatible with a background. However, existing methods always only use the real image as the positive sample to guide the training, and at most introduce the corresponding composite image as a single negative sample for an auxiliary constraint, which leads to limited distortion knowledge, and further causes a too large solution space, making the generated harmonized image distorted. Besides, none of them jointly constrain from the foreground self-style and foreground-background style consistency, which exacerbates this problem. Moreover, recent region-aware adaptive instance normalization achieves great success but only considers the global background feature distribution, making the aligned foreground feature distribution biased. To address these issues, we propose a self-consistent style contrastive learning scheme (SCS-Co). By dynamically generating multiple negative samples, our SCS-Co can learn more distortion knowledge and well regularize the generated harmonized image in the style representation space from two aspects of the foreground self-style and foreground-background style consistency, leading to a more photorealistic visual result. In addition, we propose a background-attentional adaptive instance normalization (BAIN) to achieve an attention-weighted background feature distribution according to the foreground-background feature similarity. Experiments demonstrate the superiority of our method over other state-of-the-art methods in both quantitative comparison and visual analysis.*

## 1. Introduction

Image composition is widely used in image editing [6, 45] and data augmentation [7, 46], which targets synthesizing a composite image by extracting the foreground of one image and pasting it on the background of another image. However, since the foreground and background appearance will be distinct due to different capture conditions, the composite image often looks unrealistic, *i.e.*, suffers from the inharmony problem. Therefore, image harmonization, which aims to adjust the appearance of the foreground to make it compatible with the background in the composite image, is significant and challenging.

Numerous deep learning-based methods have been proposed for image harmonization. However, most methods [6, 14, 15, 35, 39, 48] do not consider this problem from the perspective of visual style. Hence, they fail to ensure a visual style consistency between the foreground and the back-

*Equal contribution.
†Corresponding author.

ground [24]. Methods based on domain translation [4, 5] implicitly consider this problem from the perspective of domain-consistency, but do not directly transform the foreground feature in the generator.

Recently, Ling *et al*. [24] explicitly introduce the concept of visual style and first regard image harmonization as a background-to-foreground style transfer problem[1]. Inspired by AdaIN [18], they propose a region-aware adaptive instance normalization (RAIN) for image harmonization and achieve great success. However, as shown in Figure 1(d), the distortion still exists or even is very severe in some cases.

We argue that two issues lead to the above dilemma: (1) Just like the problem with AdaIN, RAIN only considers the global style distribution in the background and aligns the foreground feature distribution with it. However, as a common intuition, areas in the background that feature-similar to the foreground need more attention. For example, in the first row of Figure 1, the foreground object reappears twice in the background. The model should pay more attention to the local style distributions of these two areas. (2) The second is a general issue, not limited to the style-based method, and is the core issue we want to solve. Most existing methods [6, 14, 15, 35, 39, 48] only use real images to guide the training via an $\mathcal{L}_1$ loss, which is too simple and cannot constrain the solution space well [42]. Toward this end, DoveNet [5] and RainNet [24] adopt a domain verification loss. However, it only regards the foreground-background feature similarity of the real/harmonized image as positive/negative, and the input composite image is not used, which contains important distortion knowledge. In other words, it is just a positive-orient constraint. In addition, since image harmonization aims to adjust the foreground, why not directly constrain the foreground feature? Considering the above problems, Cong *et al*. propose a triplet loss [4]. However, it directly pulls the foreground domain code to the background domain code, which is too strong and will be interfered by content information. One more important problem is that only using the input composite image as the negative sample, leading to limited external distortion knowledge [41, 42], and the learned feature distribution easily becomes biased [25, 41]. *In summary, why not dynamically generate multiple negative samples and jointly constrain from the foreground self-style and foreground-background style consistency to obtain more distortion knowledge and reduce the solution space?*

Motivated by the observations and analyses above, we try to address these two issues. For the first issue, inspired by [26, 29], we propose a Background-attentional Adaptive Instance Normalization (BAIN). It can learn the feature similarity between the foreground and background, and calculate an attention-weighted style distribution of the background according to this feature similarity. Finally, the foreground feature distribution is aligned with this distribution.

[1]In fact, for a similar task, namely painterly harmonization, Luan *et al*. [27] introduce the concept of visual style earlier.

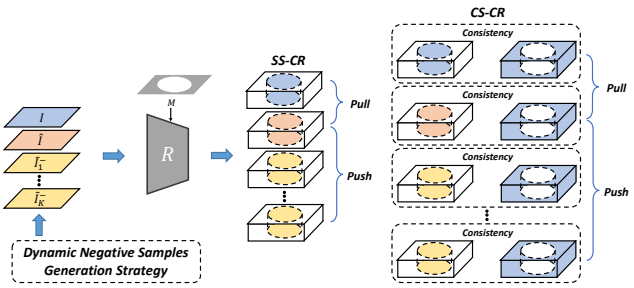



Figure 2. The illustration of our SCS-Co, including SCS-CR and the dynamic negative samples generation strategy. The detail of this strategy is shown in Figure 4.

For the second issue, we attempt to solve it by considering the positive and negative relations simultaneously in the form of contrastive learning. Specifically, we propose a novel Self-Consistent Style Contrastive Learning Scheme (SCS-Co) (see Figure 2), including a Self-Consistent Style Contrastive Regularization (SCS-CR) and a Dynamic Negative Samples Generation Strategy (see Figure 4). For a composite image $\tilde{I}$, we denote its corresponding harmonized image $\hat{I}$ and its ground truth real image $I$ as the anchor and positive sample, respectively. We also denote this composite image $\tilde{I}$ as the first negative sample $\tilde{I}_1^-$. More negative samples with the same content but different distortions are achieved via our dynamic negative samples generation strategy. Then we try to pull the anchor sample closer to the positive sample and push the anchor sample away from negative samples in the style representation space. In detail, for more powerful constraint, we not only constrain from the foreground self-style representation, but also use the background style representation as guidance to constrain from the foreground-background style consistency.

Our contributions are summarized as three-fold:

- For the first time, we introduce contrastive learning to image harmonization. Our self-consistent style contrastive learning scheme (SCS-Co) can further improve the performance of existing image harmonization networks without any increase in model parameters.
- We develop a background-attentional adaptive instance normalization (BAIN). It learns a foreground-background feature similarity attention map and properly normalizes the foreground feature by the per-point attention-weighted background feature statistics.
- Extensive experiments prove that our method is powerful for image harmonization. Compared with other state-of-the-art methods, our method obtains superior results in both quantitative metrics and visual quality.

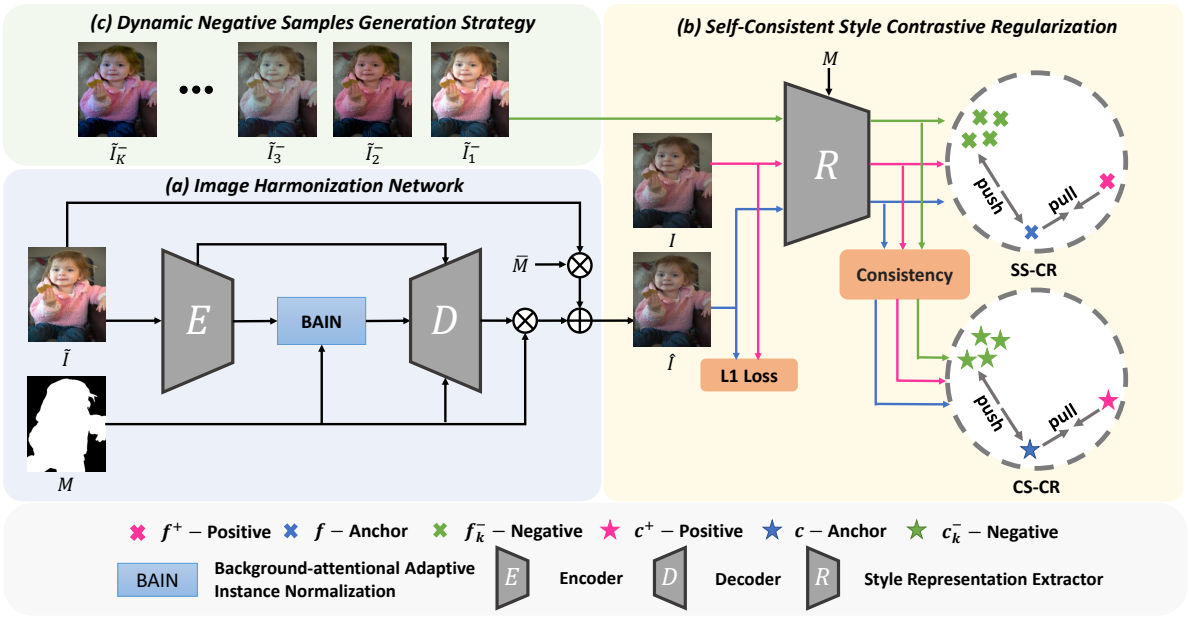



Figure 3. The architecture of our method, which consists of (a) image harmonization network, (b) self-consistent style contrastive regularization and (c) dynamic negative samples generation strategy. Please note that the first negative sample $\tilde{I}_1^-$ is the input composite image $\tilde{I}$. (b) and (c) make up our self-consistent style contrastive learning scheme.

## 2. Related Work

**Image Harmonization.** Various approaches have been proposed for image harmonization. Traditional methods focus on better transferring hand-crafted low-level appearance statistics, such as color statistics [31, 32, 45], gradient information [19, 30, 37], multi-scale various statistics [36] between foreground and background regions. However, they could not address complex cases where the foreground image has a large appearance gap with the background image. With the advances of deep learning, more deep learning-based methods were proposed. To learn the differences between various low-level features in the composite images, Cun and Pun [6] design an additional spatial-separated attention module. In [39], they present an end-to-end CNN network for image harmonization and incorporate an auxiliary segmentation branch to use semantic information. Guo *et al*. [15] first model image harmonization based on intrinsic image theory and adopt an autoencoder to disentangle composite image into reflectance and illumination for separate harmonization. In [35], they combine pre-trained semantic segmentation models with encoder-decoder architectures for image harmonization. With the rise of Transformer, Guo *et al*. [14] design the first harmonization Transformer frameworks without and with disentangled representation. In [20], they propose the first self-supervised harmonization framework that needs neither human-annotated masks nor professionally created images for training.

**Arbitrary Style Transfer.** Arbitrary style transfer is a technique used to render a photo with a particular visual style by synthesizing global and local style patterns from a given style image evenly over a content image while maintaining its original structure. Originating from non-realistic rendering [22], earlier image style transfer methods are closely related to texture synthesis [8–10]. Adopting the success of deep learning, Gatys *et al*. first formulate style transfer as the matching of multi-level deep features extracted from a pre-trained deep neural network and achieve surprising performance [11]. Huang *et al*. create a novel way for real-time style transfer by matching the mean-variance statistics between content and style features (AdaIN) [18]. Afterwards, many methods are proposed [1, 13, 26, 43, 47]. However, as stressed in [24], these style transfer methods are not practical for our task because the style defined in our work is consistent with image realism instead of texture, and our task is region-aware, which otherwise will introduce new problems of feature shift.

**Contrastive Learning.** Contrastive learning has demonstrated its effectiveness in self-supervised representation learning [3, 16, 17, 28, 33, 38, 44]. Instead of using a predefined and fixed target, contrastive learning aims to pull positive samples close to the anchor and push negative samples away in a representation space, increasing mutual information. However, different from high-level vision tasks [3, 12, 16, 17], which inherently suit for modeling the contrast between positive and negative samples, there are still few works applying contrastive learning to low-level vision tasks due to their difficulty in constructing negative samples and contrastive loss [40–42]. In this paper, specifically

for image harmonization, we design a self-consistent style contrastive learning scheme.

## 3. Our Method

### 3.1. Problem Formulation

Given a foreground image $I_f$ and a background image $I_b$, the object composition process of the composition image can be formulated as $\tilde{I} = M \cdot I_f + (1 - M) \cdot I_b$, where $\cdot$ is element-wise multiplication, $M$ is the foreground mask, which indicates the region to be harmonized, and therefore the background mask is $\bar{M} = 1 - M$. Our goal is to learn a harmonization network $G$, whose output is the harmonized image as $\hat{I} = G(\tilde{I}, M)$ and should be close to the ground truth real image $I$ by $\mathcal{L}_{rec} = \|I - \hat{I}\|_1$.

### 3.2. Image Harmonization Network

As shown in Figure 3(a), our network $G$ is based on U-Net with skip links from the encoder to the decoder. Their details can be found in supplementary. In addition, we propose a background-attentional adaptive instance normalization (BAIN) inserted between the encoder and decoder, which will be explained in detail in Section 3.4.

### 3.3. Self-Consisitent Style Contrastive Learning Scheme (SCS-Co)

As shown in Figure 2, our SCS-Co contains SCS-CR and the dynamic negative samples generation strategy. In detail, SCS-CR consists of self-style contrastive regularization (SS-CR) and consistent style contrastive regularization (CS-CR). We make it clear that self-style refers to the style of the foreground, and consistent style refers to the foreground-background style consistency.

**Formulation.** For our SCS-Co, we need to resolve two key issues. One is to construct positive and negative samples. In our SCS-Co, we choose the harmonized image $\hat{I}$ generated by the image harmonization network $G$ and the corresponding real image $I$ as the anchor and the positive sample, respectively. The most important task is to construct negative samples. We can simply use the input composite image $\tilde{I}$ as the only negative sample. However, as emphasized in existing contrastive learning methods [3, 16], a large dictionary covering a rich set of negative samples is critical for good representation learning. Therefore, during the training process, for each input composite image $\tilde{I}$, we generate $K$ negative samples online. Specifically, we propose a dynamic negative samples generation strategy. As shown in Figure 4, given an input composite image $\tilde{I}$ (Red box), we use it as the first negative sample $\tilde{I}_1^-$. Then we get its corresponding real image $I$ and segment the foreground region $R_f$ according to the foreground mask $M$. Afterwards, we sample $K - 1$ images (other than $\tilde{I}_1^-$) from the same mini-batch and segment their foreground regions. As suggested in [5, 39], we transfer the color information of these $K - 1$ foreground regions to $R_f$ respectively, leading to $K - 1$ negative samples. Finally, we successfully obtain $K$ negative samples, *i.e.*, $\tilde{I}_k^-, k = 1, 2, 3, ..., K$.

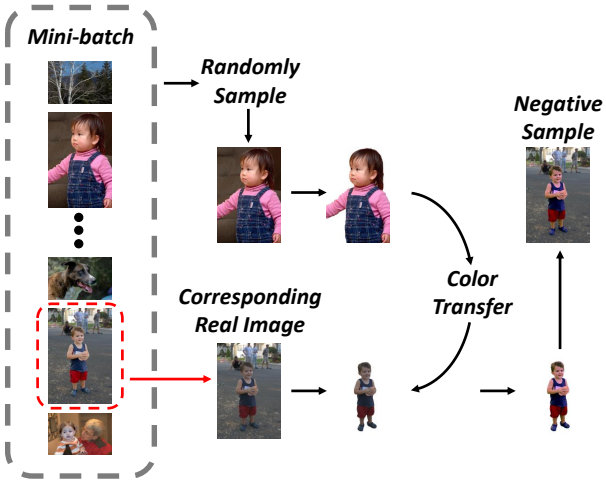



Figure 4. The illustration of our dynamic negative samples generation strategy. Red box indicates the input composite image. Through this strategy, more negative samples with the same content but different distortions are obtained, which provide much distortion knowledge.

The other is to find the style representation space of these samples for contrast. We use a fixed pre-trained style representation extractor $R$ and introduce the foreground mask $M$ and the background mask $\bar{M} = 1 - M$ to obtain style representations for different regions. Specifically, we input $\hat{I}$, $I$ and $\tilde{I}_k^-, k = 1, 2, 3, ..., K$. Then we can obtain the anchor foreground style representation $f = R(\hat{I}, M)$, the positive foreground style representation $f^+ = R(I, M)$, the positive background style representation $b^+ = R(I, \bar{M})$, and negative foreground style representations $f_k^- = R(\tilde{I}_1^-, M), k = 1, 2, 3, ..., K$.

Thus, SS-CR can be formulated as:

$$\mathcal{L}_{ss-cr} = \frac{D\left(f, f^+\right)}{D\left(f, f^+\right) + \sum_{k=1}^{K} D\left(f, f_k^-\right)}, \tag{1}$$

where $D(x, y) = \|x - y\|_1$ denotes the $\mathcal{L}_1$ distance between $x$ and $y$. As shown in Figure 2, our SS-CR focuses on the foreground self-style, pulling $f$ closer to $f^+$ and pushing $f$ away from $\{f_k^-\}_{k=1}^K$.

However, so far we have not used the background style representation, which is a powerful guidance for image harmonization [4]. Therefore, we further make contrastive constraints from the perspective of foreground-background style consistency. Specifically, we calculate the style consistency between $f$ and $b^+$ as $c = \mathrm{Gram}(f, b^+)$, where $\mathrm{Gram}(\cdot)$ means Gram Matrix [11]. Similarly, we can ob-

tain $c^+ = \text{Gram}(f^+, b^+)$, and $c_k^- = \text{Gram}(f_k^-, b^+), k = 1, 2, 3, ..., K$.

Thus, CS-CR can be formulated as:

$$\mathcal{L}_{cs-cr} = \frac{D\left(c, c^+\right)}{D\left(c, c^+\right) + \sum_{k=1}^{K} D\left(c, c_k^-\right)}, \tag{2}$$

As shown in Figure 2, our CS-CR focuses on the foreground-background style consistency, pulling $c$ closer to $c^+$ and pushing $c$ away from $\{c_k^-\}_{k=1}^K$.

Finally, the total loss function for training is:

$$\begin{aligned} \mathcal{L}_{scs-cr} &= \mathcal{L}_{ss-cr} + \mathcal{L}_{cs-cr}, \\ \mathcal{L} &= \mathcal{L}_{rec} + \lambda \cdot \mathcal{L}_{scs-cr}. \end{aligned} \tag{3}$$

where $\lambda$ is a hyperparameter for balancing the reconstruction loss and SCS-CR.

**Difference with the triplet loss.** Compared with the triplet loss [4], as shown in Figure 2, we use a contrastive learning framework. Our SCS-Co dynamically generates $K$ negative samples online and pushes the output harmonized image away from them. Through such multiple push operations, more powerful constraints can be performed in the representation space. In addition, our SCS-Co does not simply pull $f$ to $b^+$, but constrains from the perspective of foreground-background style consistency. More experiments demonstrate our SCS-Co outperforms the triplet loss for image harmonization (see Section 4.3).

### 3.4. Background-attentional Adaptive Instance Normalization (BAIN)

**Formulation.** We illustrate the structure of BAIN in Figure 5. Let $F \in \mathbb{R}^{C\times H\times W}$ be the feature map produced by the encoder and $M \in \mathbb{R}^{1\times H\times W}$ be the resized foreground mask, where $C$, $H$, $W$ indicate the number of channels, height, and width of $F$, respectively.

Specifically, in order to learn the foreground feature and background feature individually, we first separate the foreground feature map and background feature map with the corresponding mask:

$$\begin{aligned} F_b &= F \cdot \bar{M}, \\ F_f &= F \cdot M, \end{aligned} \tag{4}$$

where $F_b$ and $F_f$ are the background feature map and foreground feature map. Then, we adopt instance normalization to normalize $F_b$ and $F_f$. The normalized foreground feature map $\bar{F}_f$ at site $(c, h, w)$ is computed by:

$$\bar{F}_f^{c,h,w} = \frac{F_f^{c,h,w} - \mu_f^c}{\sigma_f^c}, \tag{5}$$

where $\mu_f^c$ and $\sigma_f^c$ denote the channel-wise mean and standard variance of the foreground feature map. Similarly, we

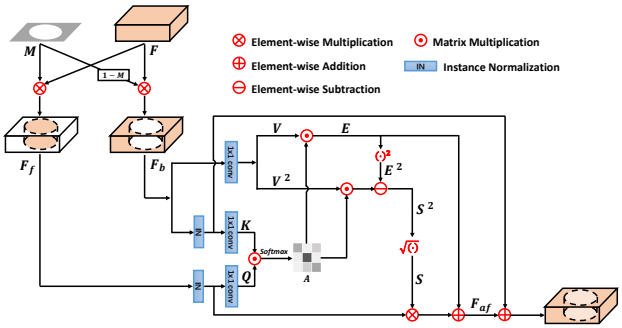



Figure 5. Background-attentional Adaptive Instance Normalization (BAIN).

can obtain the normalized background feature map $\bar{F}_b$. Further, we transform $\bar{F}_f$, $\bar{F}_b$ and $F_b$ into $Q$ (query), $K$ (key) and $V$ (value) as:

$$Q = f(\bar{F}_f), K = b(\bar{F}_b), V = k(F_b), \tag{6}$$

where $f(\cdot)$, $b(\cdot)$ and $k(\cdot)$ are $1 \times 1$ convolutions. Thus, the attention map $A \in \mathbb{R}^{HW\times HW}$ can be calculated as:

$$A = \text{Softmax}\left(Q^\top \odot K\right), \tag{7}$$

where $\odot$ indicates matrix multiplication.

Then we calculate attention-weighted background expectation and standard variance respectively. The attention-weighted background expectation $E \in \mathbb{R}^{C\times HW}$ can be calculated as:

$$E = V \odot A^\top, \tag{8}$$

Since the variance of a variable equals to the expectation of its square minus the square of its expectation, we can obtain the attention-weighted background standard variance $S \in \mathbb{R}^{C\times HW}$ as:

$$S = \sqrt{(V \cdot V) \odot A^\top - E \cdot E}, \tag{9}$$

Finally, we reshape $E$ and $S$ to $\mathbb{R}^{C\times H\times W}$, and align $\bar{F}_f$ with $E$ and $S$. The aligned foreground feature map $F_{af}$ at site $(c, h, w)$ is computed by:

$$F_{af}^{c,h,w} = S^{c,h,w} \cdot \bar{F}_f^{c,h,w} + E^{c,h,w}. \tag{10}$$

**Difference with RAIN.** Inspired by AdaIN [18], Ling *et al*. propose RAIN [24] for image harmonization and achieve great success. However, just like the problem with AdaIN, RAIN only considers the holistic style distribution in the background and globally aligns the foreground feature distribution with that of the background feature. Different from RAIN, inspired by [26, 29], our BAIN can pay more attention to those areas in the background that feature-similar to the foreground, and based on this attention map, the attention-weighted expectation and standard variance of the background feature are calculated to locally align the foreground feature distribution.

Table 1. Quantitative comparison across four sub-datasets of iHarmony4 [5]. ↑ means the higher the better, and ↓ means the lower the better. **Red** and blue indicate the best and second best performance, respectively.

| Dataset | Metric | Composite | DIH [39] | S$^2$AM [6] | DoveNet [5] | BargainNet [4] | Guo *et al*. [15] | RainNet [24] | iS$^2$AM [35] | D-HT [14] | Ours |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HCOCO | PSNR↑ | 33.94 | 34.69 | 35.47 | 35.83 | 37.03 | 37.16 | 37.08 | 39.16 | 38.76 | **39.88** |
| | MSE↓ | 69.37 | 51.85 | 41.07 | 36.72 | 24.84 | 24.92 | 29.52 | 16.48 | 16.89 | **13.58** |
| | fMSE↓ | 996.59 | 798.99 | 542.06 | 551.01 | 397.85 | 416.38 | 501.17 | 266.19 | 299.30 | **245.54** |
| HAdobe5K | PSNR↑ | 28.16 | 32.28 | 33.77 | 34.34 | 35.34 | 35.20 | 36.22 | 38.08 | 36.88 | **38.29** |
| | MSE↓ | 345.54 | 92.65 | 63.40 | 52.32 | 39.94 | 43.02 | 43.35 | 21.88 | 38.53 | **21.01** |
| | fMSE↓ | 2051.61 | 593.03 | 404.62 | 380.39 | 279.66 | 284.21 | 317.55 | 173.96 | 265.11 | **165.48** |
| HFlickr | PSNR↑ | 28.32 | 29.55 | 30.03 | 30.21 | 31.34 | 31.34 | 31.64 | 33.56 | 33.13 | **34.22** |
| | MSE↓ | 264.35 | 163.38 | 143.45 | 133.14 | 97.32 | 105.13 | 110.59 | 69.67 | 74.51 | **55.83** |
| | fMSE↓ | 1574.37 | 1099.13 | 785.65 | 827.03 | 698.40 | 716.6 | 688.40 | 443.65 | 515.45 | **393.72** |
| Hday2night | PSNR↑ | 34.01 | 34.62 | 34.50 | 35.27 | 35.67 | 35.96 | 34.83 | 37.72 | 37.10 | **37.83** |
| | MSE↓ | 109.65 | 82.34 | 76.61 | 51.95 | 50.98 | 55.53 | 57.40 | **40.59** | 53.01 | 41.75 |
| | fMSE↓ | 1409.98 | 1129.40 | 989.07 | 1075.71 | 835.63 | 797.04 | 916.48 | **590.97** | 704.42 | 606.80 |
| Average | PSNR | 31.63 | 33.41 | 34.35 | 34.76 | 35.88 | 35.90 | 36.12 | 38.19 | 37.55 | **38.75** |
| | MSE↓ | 172.47 | 76.77 | 59.67 | 52.33 | 37.82 | 38.71 | 40.29 | 24.44 | 30.30 | **21.33** |
| | fMSE↓ | 1376.42 | 773.18 | 594.67 | 532.62 | 405.23 | 400.29 | 469.60 | 264.96 | 320.78 | **248.86** |

# 4. Experiments

## 4.1. Experimental Settings

**Datasets.** Following the same setting as previous methods [5,24], we use the benchmark dataset iHarmony4 [5] to train and evaluate, which consists of four sub-datasets: HCOCO, HAdobe5k, HFlickr, and Hday2night. We follow the same settings of iHarmony4 as DoveNet [5]. We also evaluate our method on 99 real composite images released by [39].

**Evaluation Metrics.** Following [4, 5, 24, 35], the harmonized results are evaluated with Peak Signal-to-Noise Ratio (PSNR), Mean Squared Error (MSE) and foreground MSE (fMSE) on RGB channels. fMSE is an evaluation metric that only calculates MSE in the foreground region, measuring how well the foreground is harmonized.

**Compared Methods.** We compare with numerous SOTA image harmonization methods: DIH [39], DoveNet [5], RainNet [24], iS$^2$AM [35], D-HT [14], etc. We do not compare with traditional image harmonization methods since they have been proven to perform worse than deep learning methods [4, 5, 24]. All the results are either provided by the authors, or produced by their officially released codes.

**Implementation Details.** Our model is trained by Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We train the model for 120 epochs with input images resized to $256 \times 256$ and batch size set to 12. The initial learning rate is set to $10^{-3}$ and multiply by 0.1 in the 100th and 110th epoch. We use PyTorch to implement our models with Nvidia 2080Ti GPUs. Due to the powerful style representation ability of the VGG network [21], we choose the fixed pretrained VGG-16 [34] as the style representation extractor and use the latent feature at layer *relu4-3*. We set $\lambda = 0.01$ in Eq. (3). For the number of negative samples, we set $K = 5$, and will be further explored in Section 4.3.

## 4.2. Comparison with Existing Methods

**Performance on Synthesized iHarmony4 Dataset.** Table 1 shows the quantitative results of previous state-of-the-art methods as well as our method. From Table 1, we can observe that our method outperforms other compared methods on all datasets, except for MSE and fMSE value on Hday2night. Moreover, compared with the second best method, our method achieves a huge average performance gain of 0.56 dB in PSNR, 3.11 in MSE, and 16.1 in fMSE.

In Figure 6, we further present qualitative comparison results on iHarmony4. It can be easily observed that our method obtains a more consistent visual style in the whole composite image, achieving a more photorealistic output. For example, as shown in the third row of Figure 6, the visual style of the foreground and the background are quite different, resulting in obvious image distortion. The other three methods cannot adjust the style of the foreground, especially the overall tone and the contrast of lighting and shadows. Unlike them, our method produces a more photorealistic result and is closer to the ground-truth real image.

**Performance on Real Composite Images.** Figure 7 presents some results on real composite images released by DIH [39]. Because there is no ground truth image as a reference, it is impossible to compare different methods quantitatively using PSNR, MSE or fMSE. However, we can still find that our method achieves the best visual effect. Please refer to supplementary for more visual comparison.

We further conduct a user study. Following [4, 5, 15], we invite 60 volunteers and acquire a total of 29700 pairwise results for all 99 images, with 30 results for each pair of different methods on average. Then, we use the Bradley-Terry model (B-T model) [2, 23] to calculate the global ranking score for each method. Table 2 demonstrates that our method achieves the highest B-T score, which proves its effectiveness in real-world applications.


Figure 6. Qualitative comparison on samples from the testing dataset of iHarmony4 [5]. Red boxes in composite images mark foreground.

Table 2. B-T scores comparison on 99 real composite images.

| Method | Composite | DIH [39] | DoveNet [5] | RainNet [24] | Ours |
|---|---|---|---|---|---|
| B-T score↑ | 0.574 | 0.889 | 1.075 | 1.213 | **1.841** |


Figure 7. Example results on real composite images taken from [39]. Red boxes in composite images mark foreground.

### 4.3. Ablation Studies

**Effects of BAIN and SCS-Co.** In Table 3, we find that when BAIN is added, the PSNR value is improved from 37.55 dB to 37.84 dB. When SCS-Co is used, the PSNR value is improved from 37.55 dB to 38.42 dB. After adopting both of them, the PSNR value is further improved to 38.75 dB. A similar phenomenon also appears on other metrics. These comparisons demonstrate the effectiveness of our BAIN and SCS-Co, and they cooperate very well to further improve the performance. In addition, to further illustrate the effectiveness of our BAIN and SCS-Co, we show some output results of ablation experiments in Figure 8. It can be found that compared with the distortion results produced by the baseline, after adding BAIN, the color and lighting of the output results are close to the real images, but there is still a certain degree of deviation. After the introduction of SCS-Co, the deviation is further corrected, the output results are very close to the real images. More ablation studies on BAIN can be found in supplementary.

**SS-CR and CS-CR in SCS-CR.** SCS-CR is a key component of our SCS-Co and it consists of SS-CR and CS-CR. Therefore, we investigate SS-CR and CS-CR in SCS-CR. As shown in Table 4, we find that both SS-CR and CS-CR significantly improve the performance of our model, and the best result is achieved by using them all. The combination of them can strictly regularize the harmonized image in the style representation space, which significantly facilitates the generation of photorealistic visual results.

**Number of Negative Samples.** We further study the effect of the number of negative samples. As shown in Figure 9, adding more negative samples achieves better performance, because the more negative samples, the more powerful constraints can be performed. However, in Figure 9, we also observe that as the number of negative samples increases, the gain brought by adding negative samples decreases. Besides, it takes longer training time when increasing the number of negative samples. Therefore, for the performance-efficiency trade-off, we finally choose to use five negative samples, *i.e.*, we set $K = 5$.

Table 3. Performance of the baseline with BAIN and/or SCS-Co. The network with both BAIN and SCS-Co performs best.

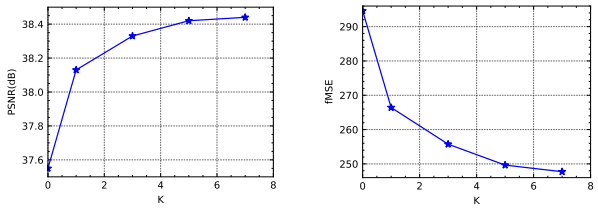
| BAIN | SCS-Co | PSNR↑ | MSE↓ | fMSE↓ |
|---|---|---|---|---|
| ✗ | ✗ | 37.55 | 27.81 | 294.64 |
| ✓ | ✗ | 37.84 | 25.23 | 269.05 |
| ✗ | ✓ | 38.42 | 22.98 | 249.65 |
| ✓ | ✓ | 38.75 | 21.33 | 248.86 |

Composite Real Baseline w/ BAIN Full model

Figure 8. Ablation study on BAIN and SCS-Co. Full model means baseline with both BAIN and SCS-Co.

Table 4. Ablation Study on SS-CR and CS-CR in SCS-CR.

| SS-CR | CS-CR | PSNR↑ | MSE↓ | fMSE↓ |
|---|---|---|---|---|
| ✗ | ✗ | 37.55 | 27.81 | 294.64 |
| ✓ | ✗ | 38.03 | 24.09 | 258.64 |
| ✗ | ✓ | 37.88 | 25.06 | 269.79 |
| ✓ | ✓ | 38.42 | 22.98 | 249.65 |




(a) PSNR (b) fMSE

Figure 9. Performance of using different numbers of negative samples in SCS-Co. We report PSNR and fMSE.

**Comparison of SCS-Co and Triplet Loss.** In section 3.3, we discuss the difference between our SCS-Co and the triplet loss in [4]. To further prove the effectiveness of our SCS-Co, we add the triplet loss to the baseline network and compare its result with our SCS-Co. In Table 5, we can find that compared with using the triplet loss, using our SCS-Co brings much more performance improvement, increasing 0.48 dB in PSNR. A similar phenomenon also appears on other metrics. Moreover, we set $K = 1$, *i.e.*, we only use the input composite image as the negative sample, which is consistent with the triplet loss. As shown in Table 5, our SCS-Co ($K = 1$) still obtains obvious improvement over the triplet loss. It proves that the improvement of our SCS-Co is not only by introducing more dynamically generated negative samples, but also by using a contrastive learning framework and constraining from the foreground self-style and foreground-background style consistency.

Table 5. Comparison of SCS-Co and the triplet loss [4].

| Method | Baseline | w/ Triplet Loss | w/ SCS-Co | w/ SCS-Co ($K = 1$) |
|---|---|---|---|---|
| PSNR↑ | 37.55 | 37.94 | 38.42 | 38.13 |
| MSE↓ | 27.81 | 25.48 | 22.98 | 23.32 |
| fMSE↓ | 294.64 | 274.65 | 249.65 | 266.43 |

Table 6. Results of integrating SCS-Co into SOTA methods.

| Method | RainNet [24] | DIH [39] | $S^2$AM [6] |
|---|---|---|---|
| PSNR↑ | 37.07(↑0.95) | 34.09(↑0.68) | 35.13(↑0.78) |
| MSE↓ | 34.92(↓5.37) | 74.72(↓2.05) | 53.86(↓5.81) |
| fMSE↓ | 364.29(↓105.31) | 707.16(↓66.02) | 538.99(↓55.68) |

**Universality of SCS-Co.** To evaluate the universality of our SCS-Co, we integrate it into three SOTA methods: RainNet [24], DIH [39] and $S^2$AM [6]. As shown in Table 6, after integrating SCS-Co, the performance of each method is improved. This proves the universality of our SCS-Co, which can be easily added to different models without any increase in model parameters.

## 5. Conclusion

In this paper, we propose a novel self-consistent style contrastive learning scheme (SCS-Co) with a self-consistent style contrastive regularization (SCS-CR) and a dynamic negative samples generation strategy. SCS-Co is built upon contrastive learning to ensure that the output harmonized image (anchor sample) is pulled closer to the real image (positive sample) and pushed away from the composite image (the first negative sample) and other dynamically generated negative samples in the style representation space. The constraint is jointly from two aspects of the foreground self-style and foreground-background style consistency. As a result, our SCS-Co can learn more distortion knowledge and reduce the solution space well. Moreover, we propose a background-attentional adaptive instance normalization (BAIN) to pay more attention to those areas in the background that feature-similar to the foreground, and the attention-weighted background feature distribution is calculated to locally align the foreground feature distribution. Experiments demonstrate that our method is superior to other SOTA methods on both synthetic and real datasets.

**Acknowledgements.** This work was partly supported by the Natural Science Foundation of Guangdong Province (No.2020A1515010711), the Natural Science Foundation of China (No.61771276) and the Special Foundation for the Development of Strategic Emerging Industries of Shenzhen (Nos.JCYJ20200109143010272 and CJGJZD20210408092804011). It is also partly supported by Overseas Cooperative Foundation.

## References


[1] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *CVPR*, pages 862–871, 2021. 3

[2] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 6

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 3, 4

[4] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Bargainnet: Background-guided domain translation for image harmonization. In *ICME*, pages 1–6, 2021. 2, 4, 5, 6, 8

[5] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *CVPR*, pages 8394–8403, 2020. 1, 2, 4, 6, 7

[6] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing*, 29:4759–4771, 2020. 1, 2, 3, 6, 8

[7] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *CVPR*, pages 1301–1310, 2017. 1

[8] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *ICCV*, volume 2, pages 1033–1038, 1999. 3

[9] Michael Elad and Peyman Milanfar. Style transfer via texture synthesis. *IEEE Transactions on Image Processing*, 26(5):2338–2351, 2017. 3

[10] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *NeurIPS*, pages 262–270, 2015. 3

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. 3, 4

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Mohammad Gheshlaghi Pires, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 3

[13] Shuyang Gu, Congliang Chen, Jing Liao, and Lu Yuan. Arbitrary style transfer with deep feature reshuffle. In *CVPR*, pages 8222–8231, 2018. 3

[14] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *ICCV*, pages 14870–14879, October 2021. 1, 2, 3, 6

[15] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *CVPR*, pages 16367–16376, 2021. 1, 2, 3, 6

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 3, 4

[17] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *ICML*, pages 4182–4192, 2020. 3

[18] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 2, 3, 5

[19] Jiaya Jia, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. Drag-and-drop pasting. *ACM Transactions on graphics (TOG)*, 25(3):631–637, 2006. 3

[20] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. Ssh: A self-supervised framework for image harmonization. In *ICCV*, pages 4832–4841, October 2021. 3

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016. 6

[22] Jan Eric Kyprianidis, John Collomosse, Tinghuai Wang, and Tobias Isenberg. State of the" art": A taxonomy of artistic stylization techniques for images and video. *IEEE Transactions on Visualization and Vomputer Graphics*, 19(5):866–885, 2012. 3

[23] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *ICCV*, pages 1701–1709, 2016. 6

[24] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *CVPR*, pages 9361–9370, 2021. 1, 2, 3, 5, 6, 7, 8

[25] Rui Liu, Yixiao Ge, Ching Lam Choi, Xiaogang Wang, and Hongsheng Li. Divco: Diverse conditional image synthesis via contrastive generative adversarial network. In *CVPR*, pages 16377–16386, 2021. 2

[26] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *ICCV*, pages 6649–6658, October 2021. 2, 3, 5

[27] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep painterly harmonization. In *Computer graphics forum*, pages 95–106, 2018. 2

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[29] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *CVPR*, pages 5880–5888, 2019. 2, 5

[30] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH*, pages 313–318, 2003. 3

[31] Francois Pitie, Anil C Kokaram, and Rozenn Dahyot. N-dimensional probability density function transfer and its application to color transfer. In *ICCV*, pages 1434–1439, 2005. 3

[32] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001. 3

[33] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *ICRA*, pages 1134–1141, 2018. 3

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[35] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *WACV*, pages 1620–1629, 2021. 1, 2, 3, 6, 7

[36] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM Transactions on Graphics*, 29(4):1–10, 2010. 3

[37] Michael W Tao, Micah K Johnson, and Sylvain Paris. Error-tolerant image compositing. In *ECCV*, pages 31–44, 2010. 3

[38] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, pages 776–794, 2020. 3

[39] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, pages 3789–3797, 2017. 1, 2, 3, 4, 6, 7, 8

[40] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *CVPR*, pages 10581–10590, 2021. 3

[41] Yanbo Wang, Shaohui Lin, Yanyun Qu, Haiyan Wu, Zhizhong Zhang, Yuan Xie, and Angela Yao. Towards compact single image super-resolution via contrastive self-distillation. *arXiv preprint arXiv:2105.11683*, 2021. 2, 3

[42] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *CVPR*, pages 10551–10560, 2021. 2, 3

[43] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *ICCV*, pages 14618–14627, October 2021. 3

[44] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018. 3

[45] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. Understanding and improving the realism of image composites. *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012. 1, 3

[46] Lingzhi Zhang, Tarmily Wen, Jie Min, Jiancong Wang, David Han, and Jianbo Shi. Learning object placement by inpainting for compositional data augmentation. In *ECCV*, pages 566–581, 2020. 1

[47] Yulun Zhang, Chen Fang, Yilin Wang, Zhaowen Wang, Zhe Lin, Yun Fu, and Jimei Yang. Multimodal style transfer via graph cuts. In *ICCV*, pages 5943–5951, 2019. 3

[48] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A Efros. Learning a discriminative model for the perception of realism in composite images. In *CVPR*, pages 3943–3951, 2015. 1, 2