BargainNet: Background-Guided Domain Translation for Image Harmonization

Wenyan Cong, Li Niu*, Jianfu Zhang, Jing Liang, Liqing Zhang
MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
{plcwyam17320,ustcnewly,c.sis,leungjing}@sjtu.edu.cn, zhang-lq@cs.sjtu.edu.cn

Abstract
Image composition is a fundamental operation in image editing field. However, unharmonious foreground and background downgrades the quality of composite image. Image harmonization, which adjusts the foreground to improve the consistency, is an essential yet challenging task. Previous deep learning based methods mainly focus on directly learning the mapping from composite image to real image, while ignoring the crucial guidance role that background plays. In this work, with the assumption that the foreground needs to be translated to the same domain as background, we formulate image harmonization task as background-guided domain translation. Therefore, we propose an image harmonization network with a novel domain code extractor and well-tailored triplet losses, which could capture the background domain information to guide the foreground harmonization. Extensive experiments on the existing image harmonization benchmark demonstrate the effectiveness of our proposed method.

1 Introduction
Image composition synthesizes the composite by combining the foreground from one image with the background from another image. One issue of image composition is the appearance differences between foreground and background caused by distinct capture conditions (e.g., weather, season, time of day). Therefore, making the generated composite realistic could be a challenging task. Image harmonization (Tsai et al. 2017; Cun and Pun 2020; Cong et al. 2020), which aims to adjust the foreground to make it compatible with the background, is essential to address this problem.

Traditional harmonization methods improve the quality of synthesized composite mainly by transferring hand-crafted appearance statistics between foreground and background regions, such as color (Cohen-Or et al. 2006; Lalonde and Efros 2007; Xue et al. 2012), texture (Sunkavalli et al. 2010), illumination (Xue et al. 2012), etc. However, they only concentrate on limited particular types of low-level features, which could not handle the large appearance gap between foreground and background regions. Recently, more deep learning based harmonization approaches have also been proposed. In (Tsai et al. 2017), they presented the first end-to-end network, which directly outputs the harmonized composites. In (Cun and Pun 2020), the spatial-separated attention blocks were proposed to learn the foreground and background features separately. Later in (Cong et al. 2020), they proposed an adversarial network with a domain verification discriminator to pull close the domains of foreground and background regions. Nonetheless, previous deep learning based methods neglected the crucial guidance role that background plays in the harmonization task. Therefore, they did not realize the shortcut to addressing image harmonization by posing it as background-guided domain translation.

According to DoveNet (Cong et al. 2020), we can treat different capture conditions as different domains. As illustrated in Figure 1(a), there could be innumerable possible domains for natural images. Even for the same scene, when the season, weather, time of the day, or photo equipment settings vary, the domain changes. For a real image, its foreground and background are captured in the same condition and thus belong to the same domain. But for a composite image, its foreground and background are likely to have different capture conditions, and thus may belong to two different domains. In this case, image harmonization could be regarded as transferring the foreground domain to the background domain, making it a special case of domain translation. Domain translation has been extensively explored in (Isola et al. 2017; Zhu et al. 2017a; Liu, Breuel, and Kautz 2017; Choi et al. 2018; Hui et al. 2018; Zhang et al. 2018; Lee et al. 2020; Choi et al. 2020), and most domain translation methods require explicitly predefined domain labels, which are unavailable in our task. More recently, methods without domain labels have also been proposed as exemplar-guided domain translation (Anokhin et al. 2020; Wang et al. 2019), which aim to translate a source image to the domain of the provided exemplar image.

In this paper, we take a further step beyond exemplar-guided domain translation. Specifically, we formulate image harmonization as a special domain translation task and detail the problem to local region guidance, i.e., background-guided domain translation. As demonstrated in Figure 1(b), the background and foreground of a composite image belong to different domains. With the guidance of extracted background domain code, which encodes the domain information of background, the composite foreground could be translated to the same domain as background, leading to a harmonious output with consistent foreground and background.

*Corresponding author.
As we propose to address image harmonization problem from a new perspective, one of our main contributions is the proposed **Background-guided domain translation Network**, which is called BargainNet for short. Since partial convolution (Liu et al. 2018) only concentrates on the feature aggregation of a partial region, we leverage partial convolution in our domain code extractor to focus on extracting the domain information of background, which can avoid the information leakage between foreground and background. The obtained background domain code defines the target domain that the foreground should be translated to, and thus helps guide the foreground domain translation. There are various ways of utilizing the target domain code to guide domain translation. For simplicity, we spatially replicate the background domain code to the same size as input image and concatenate them along the channel dimension. The concatenated input, together with the foreground mask, is fed into an attention-enhanced U-net generator (Cong et al. 2020) to produce the harmonized result. At the same time, we propose two well-tailored triplet losses to ensure that the domain code extractor can indeed extract domain information instead of domain-irrelevant information (e.g., semantic layout). The proposed triplet losses pull close the domain codes of background, real foreground, and the harmonized foreground, and push the domain code of composite foreground apart from them. To verify the effectiveness of our proposed BargainNet, we conduct comprehensive experiments on the image harmonization dataset iHarmony4 (Cong et al. 2020).

The contributions of our method are three-fold. 1) To the best of our knowledge, we are the first to formulate the image harmonization task as background-guided domain translation, which provides a new perspective for image harmonization; 2) We propose a novel image harmonization network, i.e., BargainNet, equipped with domain code extractor and well-tailored triplet losses; 3) Extensive experiments show that our image harmonization method outperforms the state-of-the-art methods.

2 Related Work

In this section, we discuss existing methods in image harmonization field. As we formulate image harmonization as background-guided domain translation, recent domain translation methods are also included in this section.

**Image Harmonization:** Image harmonization aims to make the composite foreground compatible with the background. To adjust the foreground appearance, traditional methods mainly leveraged low-level appearance statistics, such as mapping color distributions (Pitie, Kokaram, and Dahyot 2005; Reinhard et al. 2001), matching to harmonious color templates (Cohen-Or et al. 2006), applying gradient-domain compositing (Pérez, Gangnet, and Blake 2003; Jia et al. 2006; Tao, Johnson, and Paris 2010), and transferring multi-scale statistics (Sunkavalli et al. 2010). Image realism was gradually explored in (Lalonde and Efros 2007; Xue et al. 2012). In (Lalonde and Efros 2007), handcrafted color-histogram based features were used to predict visual realism and were further used to shift the foreground color distribution to improve color compatibility. In (Xue et al. 2012), they performed human realism rankings to identify key statistical measures that influence image realism and proposed to use the measures to adjust the foreground automatically.

Recently, harmonization methods that synthesize paintings from photo-realistic images have been explored in (Luan et al. 2018; Shaham, Dekel, and Michaeli 2019). However, they are more like style transfer, i.e., transferring the painting style to photo style, which is different from the photo-realistic harmonization in our task. More related to our work, in (Zhu et al. 2015), they trained a CNN model to assess and help improve the realism of composite images. In (Tsai et al. 2017), they proposed an end-to-end deep network to harmonize the input composites directly with an additional semantic parsing branch. In (Cun and Pun 2020), they inserted an attention model in the U-net structure to learn the attended feature for foreground and background separately. In (Cong et al. 2020), they proposed a GAN-structured harmonization network with a domain verification discrimina-
tor, which attempted to pull close the domains of foreground and background regions. Different from these existing methods, our proposed method provides a new perspective by treating image harmonization as a background-guided domain translation.

**Domain Translation:** The task of domain translation aims to learn the mapping from a source domain to a target domain (e.g., from day to night). Recent works (Isola et al. 2017; Zhu et al. 2017a; Liu, Breuel, and Kautz 2017; Choi et al. 2018; Hui et al. 2018; Zhang et al. 2018; Lee et al. 2020; Choi et al. 2020) have achieved impressive translation results and could be divided into two main streams depending on whether using explicit domain labels. Domain translation methods that require domain labels targeted at the conversion between two predefined domains (Isola et al. 2017; Zhu et al. 2017a; Huang et al. 2018) or between each pair of multiple domains (Choi et al. 2018; Lee et al. 2018, 2020; Liu et al. 2019). Many methods (Choi et al. 2018) focused on injecting target domain code into the conditional generator. While other methods (Liu, Breuel, and Kautz 2017; Huang et al. 2018; Lee et al. 2020) disentangled the input image into domain-invariant and domain-relevant representations, so that domain translation could be accomplished by changing domain-relevant representations. In both cases, it is still mandatory to provide explicit domain labels at both training and testing stages. There were also some works focusing on instance-level translation (Ma et al. 2018; Mo, Cho, and Shin 2019; Shen et al. 2019). Though they could achieve more fine-grained translation across domains, the domain labels are still required in these methods. In image harmonization, domains correspond to different capture conditions. Therefore, domain labels are hard to define and hard to solicit from users.

More closely related to our work, several recent approaches (Anokhin et al. 2020; Wang et al. 2019; Ma et al. 2019) without any predefined domain labels were also proposed for example-guided domain translation. Given an exemplar image as guidance, the input image is translated into the same domain as the given exemplar image. In this paper, we take a further step and pose image harmonization as background-guided domain translation, which utilizes background region instead of an exemplar image as guidance.

### 3 Our Method

In image harmonization task, we utilize training pairs of composite image $\tilde{I} \in \mathbb{R}^{H \times W \times 3}$ and real image $I \in \mathbb{R}^{H \times W \times 3}$, in which $H$ (resp., $W$) is image height (resp., width). The background of $\tilde{I}$ (real background) is the same as the background of $I$ (composite background). So in the remainder of this paper, we only mention background without distinguishing between real background and composite background. The foreground of $\tilde{I}$ (real foreground) is the harmonization target of the foreground of $I$ (composite foreground). The binary mask $M \in \mathbb{R}^{H \times W \times 1}$ indicates the foreground region to be harmonized, and therefore the complementary background mask is $\bar{M} = 1 - M$.

Given a composite image $\tilde{I}$, the goal of image harmonization task is using a generator to reconstruct $\hat{I}$ with a harmonized output $\hat{I}$, in which the foreground of $\hat{I}$ (harmonized foreground) should be close to the real foreground. Next, we first introduce our domain code extractor in Section 3.1, and then introduce our whole network BargainNet in Section 3.2.

#### 3.1 Domain Code Extractor

Intuitively, the capture condition (e.g., weather, season, time of the day) of a natural image is the same for the whole image. Therefore, after cropping an arbitrary region from a certain natural image, we can use one unified domain code to represent its domain information, which corresponds to the capture condition of this natural image.

To extract the domain code for a region with an irregular shape, our domain code extractor $E$ is composed of contiguously stacked partial convolutional layers (Liu et al. 2018), which are designed for special image generation with irregular masks. Each partial convolutional layer takes in an image/feature map and a mask to perform convolution only based on the masked region. Therefore, the output of the domain code extractor only depends on the aggregated features within the masked region, which prevents information leakage from the unmasked region. For the technical details of partial convolution, please refer to (Liu et al. 2018).

In our task, we use domain code extractor to extract the domain codes of the foreground/background regions of composite image $\tilde{I}$, real image $I$, and output image $\hat{I}$. Each of the above regions is a region from a certain natural image, so we can use one domain code to represent the domain information as aforementioned. For example, given a composite image $\tilde{I}$ and its background mask $\bar{M}$, $E$ could extract the background domain code of $\tilde{I}$. To enforce the domain code to contain domain information instead of other domain-irrelevant information (e.g., semantic layout), we use background domain code to guide the foreground domain translation and design well-tailored triplet losses to regulate the domain code, which will be introduced in the next section.

#### 3.2 Background-guided Domain Translation Network

In this section, we introduce our proposed Background-guided domain translation Network (BargainNet), which has two modules: domain code extractor $\hat{E}$ and generator $G$. The generator $G$ is attention-enhanced U-net proposed in (Cong et al. 2020) and we omit the details here.

As demonstrated in Figure 2, given a composite image $\tilde{I}$ and its background mask $\bar{M}$, the domain code extractor takes $\tilde{I}$ and $\bar{M}$ as input and outputs the background domain code $z_b$. The extracted background domain code is used as the target domain code for foreground domain translation, which means that the foreground will be translated to the background domain with its domain-irrelevant information (e.g., semantic layout) well-preserved. Besides, the background should remain unchanged if we translate it to the background domain. So for ease of implementation, we simply translate both foreground and background to the background domain. Inspired by previous domain translation methods (Zhu et al. 2017b; Choi et al. 2018), we spatially replicate the $L$-dimensional domain code $z_b$ to an
Figure 2: The network architecture of our BargainNet, which consists of attention enhanced U-Net generator $G$ and domain code extractor $E$. We employ two types of triplet losses based on four types of domain codes (see Section 3.2). The test phase is highlighted with red flow lines for clarity.

$H \times W \times L$ domain code map $Z_0$ and concatenate it with the $H \times W \times 3$ composite image. Besides, based on our experimental observation (see Section 4.4), it is still necessary to use foreground mask to indicate the foreground region to be harmonized as in previous harmonization methods (Tsai et al. 2017; Cun and Pun 2020; Cong et al. 2020), probably because the foreground mask emphasizes foreground translation and enables the foreground to borrow information from the background. Thus, we further concatenate the input with the $H \times W \times 1$ foreground mask $M$, leading to the final $H \times W \times (L + 4)$ input. After passing the input through the generator $G$, we enforce the harmonized output $\hat{I} = G(I, M, Z_{fb})$ to be close to the ground-truth real image $I$ by using the reconstruction loss $\mathcal{L}_{\text{rec}} = \|I - \hat{I}\|_1$.

By using the target domain code $z_b$ to guide foreground domain translation, we assume that $z_b$ only contains the domain information of background. Because if $z_b$ contains the domain-irrelevant information (e.g., semantic layout) of background, it may corrupt the semantic layout of foreground, which violates the reconstruction loss. To further reinforce our assumption on domain code, we use triplet losses to pull close the domain codes which are expected to be similar and push apart those which are expected to be divergent. Analogous to extracting background domain code $z_b$, we also use $\hat{E}$ to extract the domain codes of real foreground, composite foreground, and harmonized foreground, which are denoted as $z_f$, $\hat{z}_f$, and $\hat{z}_f$ respectively.

Hence, the domain code of harmonized foreground ($\hat{z}_f$) should be close to that of background ($z_b$), but far away from that of composite foreground ($\hat{z}_f$). For ease of description, we define an image triplet as a composite image, its ground-truth real image, and its harmonized output. Given an image triplet, we can obtain $\hat{z}_f$, $z_b$, and $\hat{z}_f$. Following the concepts (anchor, positive, and negative) of triplet loss, we treat $\hat{z}_f$ (resp., $z_b$, $\hat{z}_f$) as anchor (resp., positive, negative). In other words, we aim to pull close $\hat{z}_f$ and $z_b$ while pushing apart $\hat{z}_f$ and $\hat{z}_f$, which can be achieved by the following triplet loss:

$$\mathcal{L}_{f_b} = \mathcal{L}(\hat{z}_f, z_b, \hat{z}_f) = \max(d(\hat{z}_f, z_b) - d(\hat{z}_f, \hat{z}_f) + m, 0),$$

in which $d(\cdot, \cdot)$ is Euclidean distance and $m$ is a margin. By minimizing $\mathcal{L}_{f_b}$, we expect $d(\hat{z}_f, z_b)$ to be smaller than $d(\hat{z}_f, \hat{z}_f)$ by at least a margin $m$.

Next, we consider the relationship among three foregrounds in an image triplet. The harmonized foreground should belong to the same domain as the real foreground. This common domain shared by harmonized foreground and real foreground should be different from that of the composite foreground. Therefore, we enforce $d(\hat{z}_f, \hat{z}_f)$ to be smaller than $d(z_f, \hat{z}_f)$ also by at least a margin $m$, which can be achieved by the following triplet loss:

$$\mathcal{L}_{f_f} = \mathcal{L}(z_f, \hat{z}_f, \hat{z}_f) = \max(d(z_f, \hat{z}_f) - d(z_f, \hat{z}_f) + m, 0).$$

In fact, there could be many reasonable combinations of triplet losses to regulate the domain code. However, based
Table 1: Quantitative comparison between our proposed BargainNet and other baseline methods. The best results are denoted in boldface.

| Sub-dataset | HCOCO | HAdobe5k | HFlickr | Hday2night | All |
|-------------|-------|----------|---------|------------|-----|
| **Evaluation metric** | **MSE** | **PSNR** | **MSE** | **PSNR** | **MSE** | **PSNR** | **MSE** | **PSNR** | **MSE** | **PSNR** |
| Input composite | 69.37 | 33.94 | 345.54 | 28.16 | 264.35 | 28.32 | 109.65 | 34.01 | 174.27 | 31.63 |
| Lalonde and Efros(2007) | 110.10 | 31.14 | 158.90 | 29.66 | 329.87 | 26.43 | 199.93 | 29.80 | 150.53 | 30.16 |
| Xue et al.(2012) | 77.04 | 33.32 | 274.15 | 28.79 | 249.54 | 28.32 | 190.51 | 31.24 | 155.87 | 31.40 |
| Zhu et al.(2015) | 79.82 | 33.04 | 414.31 | 27.26 | 315.42 | 27.52 | 136.71 | 32.32 | 204.77 | 30.72 |
| DIH (2017) | 51.85 | 34.69 | 92.65 | 32.28 | 163.38 | 29.55 | 82.34 | 34.62 | 76.77 | 33.41 |
| $S^2$AM (2020) | 41.07 | 35.47 | 63.40 | 33.77 | 143.45 | 30.03 | 76.61 | 34.50 | 59.67 | 34.35 |
| DoveNet (2020) | 36.72 | 35.83 | 52.32 | 34.34 | 133.14 | 30.21 | 54.05 | 35.18 | 52.36 | 34.75 |
| **Ours** | **24.84** | **37.03** | **39.94** | **35.34** | **97.32** | **31.34** | **50.98** | **35.67** | **37.82** | **35.88** |

Table 2: MSE and foreground MSE (fMSE) of different methods in each foreground ratio range based on the whole test set. The best results are denoted in boldface.

| Foreground ratios | 0% ~ 5% | 5% ~ 15% | 15% ~ 100% | 0% ~ 100% |
|-------------------|---------|----------|-----------|----------|
| **Evaluation metric** | **MSE** | **IMSE** | **MSE** | **IMSE** | **MSE** | **IMSE** | **MSE** | **IMSE** |
| Input composite | 28.51 | 1208.86 | 119.19 | 1323.23 | 577.58 | 1887.05 | 172.47 | 1387.30 |
| Lalonde and Efros(2007) | 41.52 | 1481.59 | 120.62 | 1309.79 | 444.65 | 1467.98 | 150.53 | 1433.21 |
| Xue et al.(2012) | 31.24 | 1325.96 | 132.12 | 1459.28 | 479.53 | 1555.69 | 155.87 | 1411.40 |
| Zhu et al.(2015) | 33.30 | 1297.65 | 145.14 | 1577.70 | 682.69 | 2251.76 | 204.77 | 1580.17 |
| DIH (2017) | 18.92 | 799.17 | 64.23 | 725.86 | 228.86 | 768.89 | 76.77 | 773.18 |
| $S^2$AM (2020) | 15.09 | 623.11 | 48.33 | 540.54 | 177.62 | 592.83 | 59.67 | 594.67 |
| DoveNet (2020) | 14.03 | 591.88 | 44.90 | 504.42 | 152.07 | 505.82 | 52.36 | 549.96 |
| **Ours** | **10.55** | **450.33** | **32.13** | **359.49** | **109.23** | **353.84** | **37.82** | **405.23** |

4 Experiments

4.1 Dataset

We evaluate our method and baselines on the benchmark dataset iHarmony4 (Cong et al. 2020), which contains 73146 pairs of synthesized composite images and the ground-truth real images (65742 pairs for training and 7404 pairs for testing). iHarmony4 consists of four sub-datasets: HCOCO, HAdobe5k, HFlickr, and Hday2night. The details of four sub-datasets can be found in the Supplementary.

4.2 Implementation Details

The domain code extractor $E$ is formed by seven partial convolutional layers with kernel size 3 and stride 2, each of which is followed by ReLU and batch normalization except the last one. The extracted domain code is a 16-dimension vector. We set the margin $m$ in Eqn. (1)(2) as 1 and the trade-off parameter $\lambda$ in Eqn. (3) as 0.01. In our experiments, the input images are resized to $256 \times 256$ during both training and testing phases. Following (Tsai et al. 2017; Cong et al. 2020), we use Mean-Squared Errors (MSE) and Peak Signal-to-Noise Ratio (PSNR) as the main evaluation metrics, which are also calculated on $256 \times 256$ images. More details can be found in Supplementary.

4.3 Comparison with Existing Methods

Both traditional methods (Lalonde and Efros 2007; Xue et al. 2012) and deep learning based methods (Zhu et al. 2015; Tsai et al. 2017; Cun and Pun 2020; Cong et al. 2020) are included for quantitative comparisons. Following (Tsai et al. 2017; Cong et al. 2020), we train the model on the merged training sets of four sub-datasets in iHarmony4. The trained model is evaluated on each test set and the merged test set as well. Table 1 shows the quantitative results of different harmonization methods and the results of previous baselines are directly copied from (Cong et al. 2020).

From Table 1, we can observe that our method not only significantly exceeds traditional methods, but also outperforms deep learning based approaches on all sub-datasets. Besides, following DoveNet (Cong et al. 2020), we also report the MSE and foreground MSE (fMSE) on the test images in different foreground ratio ranges (e.g., 5% ~ 15%). The foreground ratio means the area of the foreground over the area of the whole image. Foreground MSE (fMSE) is MSE calculated only in the foreground region. As shown in Table 2, our method outperforms all the baselines in each foreground ratio range, which demonstrates the robustness of our proposed method.

Signal-to-Noise Ratio (PSNR) as the main evaluation metrics, which are also calculated on $256 \times 256$ images. More details can be found in Supplementary.
Table 3: The ratio of training/testing image triplets which satisfy the specified requirements. Note that $d_{x,y}$ is short for $d(z_x, z_y)$. For example, $d_{b,f}$ denotes the Euclidean distance between the background domain code $z_b$ and the domain code of real foreground $z_f$.

| # | mask | $z_b$ | $\mathcal{L}_{f,f}$ | $\mathcal{L}_{f,b}$ | MSE ↓ | PSNR ↑ |
|---|------|------|-----------------|-----------------|------|------|
| 1 | ✓ | ✓ | 60.79 | 34.15 |
| 2 | ✓ | ✓ | 43.70 | 35.43 |
| 3 | ✓ | ✓ | ✓ | 37.82 | 35.88 |
| 4 | ✓ | ✓ | ✓ | 41.03 | 35.47 |
| 5 | ✓ | ✓ | ✓ | 41.71 | 35.50 |
| 6 | ✓ | ✓ | ✓ | 115.48 | 31.94 |
| 7 | ✓ | ✓ | ✓ | 120.49 | 31.89 |

Table 4: Ablation studies on input format and triplet losses. “mask” means foreground mask and $z_b$ denotes the background domain code. Two triplet losses are $\mathcal{L}_{f,f}$ and $\mathcal{L}_{f,b}$.

4.4 Ablation Studies

As described in Section 3.2, we concatenate the composite image, foreground mask, and background domain code map as input for our generator $G$. Now we investigate the impact of each type of input and report the results in Table 4. When we only use composite image and foreground mask as input (row 1), it is exactly the same as the attention-enhanced U-net introduced in (Cong et al. 2020). After adding the background domain code to the input (row 2), the performance is significantly boosted, which demonstrates that background domain code can provide useful guidance for foreground harmonization. We further apply our proposed two triplet losses to regulate the domain code (row 3), which brings in extra performance gain. This is because that the triplet losses impose reasonable constraints for better domain code extraction. In addition, we also investigate the case in which we only feed the generator with composite image and the background domain code map while removing the foreground mask from input. No matter using triplet losses (row 6) or not (row 7), the performance is significantly degraded after removing the foreground mask (row 6 v.s. row 3, row 7 v.s. row 2), probably because the foreground mask emphasizes foreground translation and enables the foreground to borrow information from background.

Besides, we also ablate each type of triplet loss (row 4 and row 5) in Table 4. The results demonstrate that each type of triplet loss is helpful, and two types of triplet losses can collaborate with each other to achieve further improvement.

4.5 Domain Code Analyses

Recall that we employ two triplet losses Eqn. (1)(2) to regulate the domain code. To verify that the expected requirements are satisfied on the training set and generalizable to the test set, we conduct domain code analyses on both training set and test set. As defined in Section 3.2, an image triplet contains a composite image, its ground-truth real image, and its harmonized output. We calculate the ratio of training/testing image triplets which satisfy $d(\hat{z}_f, z_f) < d(\hat{z}_f, \hat{z}_f)$ (resp., $d(\hat{z}_f, z_f) < d(z_f, \hat{z}_f)$) corresponding to Eqn. (1) (resp., Eqn. (2)). For brevity, we use $d_{x,y}$ to denote $d(z_x, z_y)$, as shown in Table 3.

More generally, in an image triplet, the background, the real foreground, and the harmonized foreground belong to the same domain, while the composite foreground belongs to another domain. Considering that the distance between cross-domain regions should be larger than the distance between same-domain regions, we could construct 6 groups of (anchor, positive, negative) in the form of triplet loss, leading to 6 requirements: $d_{b,f} < d_{b,j}, d_{b,f} < d_{b,j}, d_{f,f} < d_{f,j}, d_{f,f} < d_{f,j}, d_{f,b} < d_{f,b}, d_{f,b} < d_{f,j}$. The verification results of each individual requirement and all requirements are summarized in Table 3. We can observe the high ratio of training/testing image triplets that satisfy each individual requirement. Moreover, most training/testing image triplets satisfy all six requirements at the same time. This implies that the domain code extractor can indeed extract the domain code which contains domain information as expected.

4.6 Hyper-parameter Analyses

We investigate the impact of three hyper-parameters: the margin $m$ in Eqn. (1)(2), $\lambda$ in Eqn. (3) and the domain code dimension $L$. In Figure 3, we plot the performance by varying each hyper-parameter while keeping the other hyper-parameters fixed. It can be seen that our method is robust with $m$ (resp., $\lambda$) in a reasonable range $[2^{-2}, 2^2]$ (resp., $[10^{-4}, 10^{-1}]$). With the domain code dimension increasing to 16, the performance improves obviously. When $L$ is larger than 16, the performance increases marginally, but more training resources are in demand. So 16-dimensional domain code is a cost-effective choice.
4.7 Qualitative Analyses

Given an input composite image from the test set, the harmonized outputs generated by DIH (2017), S²AM (2020), DoveNet (2020), BargainNet (w/o $\mathcal{L}_{tri}$) and BargainNet are shown in Figure 4. BargainNet (w/o $\mathcal{L}_{tri}$) is a special case without triplet losses. Compared with other baselines, BargainNet could generate more favorable results with consistent foreground and background, which are visually closer to the ground-truth real images. Besides, by comparing BargainNet with BargainNet (w/o $\mathcal{L}_{tri}$), we can observe that the generated outputs of BargainNet are more harmonious after using triplet losses, which provides an intuitive demonstration that triplet losses contribute to more effective domain code extraction.

In the real-world applications, given a real composite image, there is no ground-truth as the synthesized composite, so it is infeasible to evaluate the model performance quantitatively using MSE or PSNR. Following (Tsai et al. 2017; Cun and Pun 2020; Cong et al. 2020), we conduct user study on 99 real composite images (Tsai et al. 2017), in which we compare our BargainNet with all the other deep learning based methods. The details of user study and the global ranking results can be found in the Supplementary, from which we can observe that our BargainNet shows an advantage over the other methods with the highest B-T score. For more visual comparison on the real composite images, please refer to the Supplementary as well.

4.8 Background Harmonization and Inharmony Level Prediction

By inverting the mask fed into the generator and the domain code extractor in the testing stage, our BargainNet could be easily applied to background harmonization. We show our background harmonization results and compare with other deep learning based methods in Supplementary.

Besides, one byproduct of our method is predicting the inharmony level of a composite image, which reflects how inharmonious this composite image is. In particular, based on the extracted domain codes of the foreground region and background region, we can assess the inharmony level by calculating the Euclidean distance between two domain codes. The detailed inharmony level analyses are also left to Supplementary due to space limitation.

5 Conclusion

In this work, we have proposed to formulate image harmonization as background-guided domain translation. It is the first attempt to address image harmonization from such perspective. We have also presented BargainNet, a novel network that leverages the background domain code for foreground harmonization. Experimental results have shown that our method performs favorably on both the synthesized dataset iHarmony4 and real composite images against other state-of-the-art methods.

References

Anokhin, I.; Solovey, P.; Korzenkov, D.; Kharlamov, A.; Khakhulin, T.; Silvestrov, A.; Nikolenko, S.; Lempitsky, V.; and
Zhu, J.; Park, T.; Isola, P.; and Efros, A. A. 2017a. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2242–2251.

Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; and Shechtman, E. 2017b. Toward multimodal image-to-image translation. In Advances in Neural Information Processing Systems (NeurIPS), 465–476.
Supplementary Material for BargainNet: Background-Guided Domain Translation for Image Harmonization

Wenyan Cong, Li Niu*, Jianfu Zhang, Jing Liang, Liqing Zhang
MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
{plcwyam17320,ustcnewly,c.sis,leungjing}@sjtu.edu.cn, zhang-lq@cs.sjtu.edu.cn

In this Supplementary file, we will introduce the details of iHarmony4 dataset and our network implementation infrastructure in Section 1, 2. Then, we will show the significance test between the strongest baseline DoveNet and our BargainNet in Section 3. Besides, we will introduce more details of user study conducted on real composite images and show some harmonized results of deep learning based methods on real composite images in Section 4. Finally, we will exhibit the background harmonization results of deep learning based methods in Section 5, and analyze the inharmony level prediction of our method in Section 6.

1 Dataset Statistics

The iHarmony4 dataset contributed by (Cong et al. 2020) is composed of pairs of synthesized composite images and the ground-truth real images. iHarmony4 consists of 4 sub-datasets: HCOCO, HAdobe5k, HFlickr, and Hday2night.

HCOCO sub-dataset is synthesized based on the merged training and test splits of Microsoft COCO (Lin et al. 2014), containing 38545 training and 4283 test pairs of composite and real images. In HCOCO, the composite images are synthesized from real images and the foreground of composite image is adjusted by transferring the color from another foreground object of the same class in COCO using color mapping functions.

HAdobe5k sub-dataset is generated based on MIT-Adobe FiveK dataset (Bychkovsky et al. 2011), containing 19437 training and 2160 test pairs of composite and real images. The composite image is generated by exchanging the manually segmented foreground between the real image and its five different renditions.

HFlickr sub-dataset is synthesized based on the crawled images from Flickr, containing 7449 training and 828 test pairs of composite and real images. The composite images are synthesized similarly to HCOCO, except that the reference foreground is selected from ADE20K (Zhou et al. 2019) using the dominant category labels generated by pretrained scene parsing model.

Hday2night sub-dataset is generated based on day2night (Zhou, Sattler, and Jacobs 2016), containing 311 training

*Corresponding author.

Table 1: The results (Mean±Std) of 10 runs for our method and DoveNet.

| Method       | MSE  | PSNR  |
|--------------|------|-------|
| DoveNet      | 52.76±1.38 | 34.72±0.16 |
| BargainNet   | 38.19±0.40 | 35.83±0.06 |

2 Implementation Details

Our network is trained on ubuntu 16.04 LTS operation system, with 64GB memory, Intel Core i7-8700K CPU, and two GeForce GTX 1080 Ti GPUs. The network is implemented using Pytorch 1.4.0 and the weight is initialized with values drawn from the normal distribution $N(\text{mean} = 0.0, \text{std}^2 = 0.02)$.

3 Significance Test

To better demonstrate the robustness of our method, we train the strongest baseline DoveNet and our BargainNet each for ten times on iHarmony4 dataset and compare their performance based on the mean and variance of MSE and PSNR. The MSE (resp. PSNR) result of our method on the whole test set is Mean±Std=38.19±0.40 (resp. Mean±Std=35.83±0.06), while the MSE (resp. PSNR) result of DoveNet is Mean±Std=52.76±1.38 (resp. Mean±Std=34.72±0.16). At the significance level 0.05, we conduct significance test to verify the superiority of our BargainNet. The p-values w.r.t. MSE/PSNR are $7.55e^{-12}/2.10e^{-10}$ on the whole test set, which are far below the specified significance level. This is strong evidence that our method is significantly better than the state-of-the-art DoveNet.

4 Results on Real Composite Images

In reality, there is no ground-truth real image for a given composite image, so it is infeasible to evaluate model performance quantitatively in such a scenario. Therefore, follow-
Figure 1: Example results of background harmonization. From left to right, we show the input composite image, ground-truth real image for foreground harmonization, the background harmonization results of DIH (2017), S\(^2\)AM (2020), DoveNet (2020), and our proposed BargainNet. For clarity, we highlight the foreground with red border lines.

| Method      | B-T score↑ |
|-------------|------------|
| Input composite | 0.431      |
| DIH (2017)   | 0.820      |
| S\(^2\)AM (2020) | 0.902      |
| DoveNet (2020) | 1.145      |
| Ours         | 1.202      |

Table 2: B-T scores of deep learning based methods on 99 real composite images provided in (Tsai et al. 2017).

Similarly, given each composite image and its four harmonized outputs from four different methods, we can construct image pairs \((I_i, I_j)\) by randomly selecting from these five images \(\{I_i^k | k=1\}\). Hence, we can construct a large number of image pairs based on 99 real composite images. Each user involved in this subjective evaluation could see an image pair each time to decide which one looks more realistic. Considering the user bias, 22 users participate in the study in total, contributing 10835 pairwise results. With all pairwise results, we employ the Bradley-Terry model (B-T model) (Bradley and Terry 1952; Lai et al. 2016) to obtain the global ranking of all methods and the results are reported in Table 2. Our proposed BargainNet shows an advantage over other deep-based methods with the highest B-T score, which demonstrates that by explicitly using the background domain code as guidance, our method could generate more favorable results in real-world applications.

In Figure 3, we present some results from real composite images used in our user study. We compare the real composite images with harmonization results generated by our proposed method and other deep learning based methods, including DIH (Tsai et al. 2017), S\(^2\)AM (Cun and Pun 2020), and DoveNet (Cong et al. 2020). Based on Figure 3, we can see that our proposed method could generally produce satisfactory harmonized images compared to other deep learning based methods.

5 Generalization to Background Harmonization

Interestingly, our method could also be used for background harmonization, which means adjusting the background to make it compatible with the foreground. In particular, we can feed the composite image \(\tilde{I}\), the background mask \(\tilde{M}\), and the composite foreground domain code \(\tilde{z}_f\) into our gen-
Figure 2: Examples of composite images with different inharmony levels. From top to bottom, we show the network input and the harmonized output of our BargainNet respectively. From left to right, we show the five composite images and the ground-truth real image. The number below each image is its inharmony score.

6 Inharmony Level Prediction

Based on the extracted domain codes of foreground and background, we can predict the inharmony level of a composite image, reflecting to which extent the foreground is incompatible with the background.

We conduct experiments on HAdobe5k sub-dataset, because each real image in MIT-Adobe FiveK dataset (Bychkovsky et al. 2011) has another five edited renditions of different styles. Given a real image, we can paste the foregrounds of five edited renditions on the background of real image, leading to five composite images with the same background yet different foregrounds in HAdobe5k. Therefore, when feeding the five composite images into $G$, the generated outputs are expected to be harmonized to the same ground-truth real image. Recall that in our BargainNet, we propose to use domain code extractor to extract the domain codes $\tilde{z}_f$ and $z_b$ for foreground and background respectively. So we can calculate the Euclidean distance $d(z_b, \tilde{z}_f)$ as the inharmony score of a composite image, which reflects how inharmonious a composite image is. For the composite images with high inharmony scores, the foreground and the background are obviously inconsistent. After harmonization, the composite foreground is adjusted to be compatible with the background. Therefore, the inharmony score should become lower. In Figure 2, we show one ground-truth real image with its five composite images from HAdobe5k sub-dataset, and report two inharmony scores of each image before and after harmonization. In the top row, we can observe that composite images whose foreground and background are obviously inconsistent have higher scores, while the ground-truth real image with consistent foreground and background has the lowest inharmony score. In the bottom row, after harmonization, as the foreground is translated to the same domain as background, the inharmony score of the harmonized output decreases dramatically. Interestingly, even for the ground-truth real image, harmonization using our method can further lower its inharmony score, probably because our network could make the foreground domain closer to the background.

Inharmony level provides an intuitive perspective for inharmony assessment, which is an enticing byproduct of our method and useful for harmonization related tasks. For example, given abundant composite images, we can first predict their inharmony levels and only harmonize those with high inharmony levels for computational efficiency.

References

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39(3/4): 324–345.

Bychkovsky, V.; Paris, S.; Chan, E.; and Durand, F. 2011. Learning photographic global tonal adjustment with a database of input / output image pairs. In *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 97–104.

Cong, W.; Zhang, J.; Niu, L.; Liu, L.; Ling, Z.; Li, W.; and Zhang, L. 2020. DoveNet: Deep Image Harmonization via Domain Verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8391–8400.
Figure 3: The harmonized results on real composite images, including three deep learning based methods and our proposed BargainNet.
Cun, X.; and Pun, C. 2020. Improving the Harmony of the Composite Image by Spatial-Separated Attention Module. *IEEE Trans. Image Process.* 29: 4759–4771.

Lai, W.; Huang, J.; Hu, Z.; Ahuja, N.; and Yang, M. 2016. A Comparative Study for Single Image Blind Deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1701–1709.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, 740–755.

Tsai, Y.; Shen, X.; Lin, Z.; Sunkavalli, K.; Lu, X.; and Yang, M. 2017. Deep Image Harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2799–2807.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic Understanding of Scenes Through the ADE20K Dataset. *Int. J. Comput. Vis.* 127(3): 302–321.

Zhou, H.; Sattler, T.; and Jacobs, D. W. 2016. Evaluating Local Features for Day-Night Matching. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, 724–736.