DoveNet: Deep Image Harmonization via Domain Verification

Wenyan Cong[1], Jianfu Zhang[1], Li Niu[1]*, Liu Liu[1], Zhixin Ling[1], Weiyuan Li[2], Liqing Zhang[1]
[1] MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University [2] East China Normal University
[1]{plcwyam17320,c.sis,ustcnewly,Shirlley,1069066484}@sjtu.edu.cn
[2]10162100162@stu.ecnu.edu.cn [1]zhang-lq@cs.sjtu.edu.cn

Abstract

Image composition is an important operation in image processing, but the inconsistency between foreground and background significantly degrades the quality of composite image. Image harmonization, aiming to make the foreground compatible with the background, is a promising yet challenging task. However, the lack of high-quality publicly available dataset for image harmonization greatly hinders the development of image harmonization techniques. In this work, we contribute an image harmonization dataset iHarmony4 by generating synthesized composite images based on COCO (resp., Adobe5k, Flickr, day2night) dataset, leading to our HCOCO (resp., HAdobe5k, HFlickr, Hday2night) sub-dataset. Moreover, we propose a new deep image harmonization method DoveNet using a novel domain verification discriminator, with the insight that the foreground needs to be translated to the same domain as background. Extensive experiments on our constructed dataset demonstrate the effectiveness of our proposed method. Our dataset and code are available at https://github.com/bcmi/Image_Harmonization_Datasets.

1. Introduction

Image composition targets at generating a composite image by extracting the foreground of one image and pasting it on the background of another image. However, since the foreground is usually not compatible with the background, the quality of composite image would be significantly downgraded. To address this issue, image harmonization aims to adjust the foreground to make it compatible with the background in the composite image. Both traditional methods [20, 47, 54] and deep learning based method [43, 45] have been explored for image harmonization, in which deep learning based method [43, 45] could achieve promising results.

As a data-hungry approach, deep learning calls for a large number of training pairs of composite image and harmonized image as input image and its ground-truth output. However, given a composite image, manually creating its harmonized image, *i.e.*, adjusting the foreground to be compatible with background, is in high demand for extensive efforts of skilled expertise. So this strategy of constructing datasets is very time-consuming and expensive, making it infeasible to generate large-scale training data. Alternatively, as proposed in [43], we can treat a real image as harmonized image, segment a foreground region, and adjust this foreground region to be inconsistent with the background, yielding a synthesized composite image. Then, pairs of synthesized composite image and real image can be used to supersede pairs of composite image and harmonized image. Because foreground adjustment can be done automatically (*e.g.*, color transfer methods) without time-consuming expertise editing, it becomes feasible to collect large-scale training data. Despite this inspiring strategy proposed in [43], Tsai *et al.* [43] did not make the constructed datasets publicly available. Besides, the proposed dataset has several shortcomings, such as inadequate diversity/realism of synthesized composite images and lack of real composite images.

Considering the unavailability and shortcomings of the dataset built in [43], we tend to build our own stronger dataset. Overall, we adopt the strategy in [43] to generate pairs of synthesized composite image and real image. Similar to [43], we generate synthesized composite images based on Microsoft COCO dataset [24], MIT-Adobe5k dataset [2], and our self-collected Flickr dataset. For Flickr dataset, we crawl images from Flickr image website by using the category names in ImageNet dataset [5] as queries in order to increase the diversity of crawled images. Nevertheless, not all crawled images are suitable for the image harmonization task. So we manually filter out the images with pure-color or blurry background, the cluttered images with no obvious foreground objects, and the images which appear apparently unrealistic due to artistic editing.

Besides COCO, Adobe5k, and Flickr suggested in [43], we additionally consider datasets which contain multiple

*Corresponding author.

images captured in different conditions for the same scene. Such datasets are naturally beneficial for image harmonization task because composite images can be easily generated by replacing the foreground region in one image with the same foreground region in another image. More importantly, two foreground regions are both from real images and thus the composite image is actually a real composite image. However, to the best of our knowledge, there are only a few available datasets [40, 53, 18] within this scope. Finally, we choose day2night dataset [18], because day2night provides a collection of aligned images captured in a variety of conditions (*e.g.*, weather, season, time of day) for each scene. According to the names of original datasets, we refer to our constructed sub-datasets as HCOCO, HAdobe5k, HFlickr, and Hday2night, with "H" standing for "Harmonization". All four sub-datasets comprise a large-scale image harmonization dataset. The details of constructing four sub-datasets and the difference from [43] will be fully described in Section 3.

As another contribution, we propose DoveNet, a new deep image harmonization method with a novel domain verification discriminator. Given a composite image, its foreground and background are likely to be captured in different conditions (*e.g.*, weather, season, time of day), and thus have distinctive color and illumination characteristics, which make them look incompatible. Following the terminology in domain adaptation [32, 29] and domain generalization [31, 30], we refer to each capture condition as one domain and there could be numerous possible domains. In this case, the foreground and background of a composite image belong to two different domains, while the foreground and background of a real image belong to the same domain. Therefore, the goal of image harmonization, *i.e.*, adjusting the foreground to be consistent with background, can be deemed as translating the domain of foreground to the same one as background without knowing the domain labels of foreground and background. Inspired by adversarial learning [9, 11], we propose a domain verification discriminator to pull close the domains of foreground and background in a harmonized image. Specifically, we treat the paired foreground and background representations of a real (*resp.*, composite) image as a positive (*resp.*, negative) pair. On the one hand, we train the discriminator to distinguish positive pairs from negative pairs. On the other hand, the generator is expected to produce a harmonized image, which can fool the discriminator into perceiving its foreground-background pair as positive. To verify the effectiveness of our proposed domain verification discriminator, we conduct comprehensive experiments on our constructed dataset. Our main contributions are summarized as follows:

- We release the first large-scale image harmonization dataset iHarmony4 consisting of four sub-datasets: HCOCO, HAdobe5K, HFlickr, and Hday2night.
- We are the first to introduce the concept of domain verification, and propose a new image harmonization method DoveNet equipped with a novel domain verification discriminator.

2. Related Work

In this section, we review the development of image harmonization. Besides, as image harmonization is a special case of image-to-image translation, we discuss other related applications in this realm.

Image Harmonization: Traditional image harmonization methods concentrated on better matching low-level appearance statistics, such as matching global and local color distributions [35, 37], mapping to predefined perceptually harmonious color templates [4], applying gradient-domain compositing [34, 14, 42], and transferring multi-scale various statistics [41]. To link lower-level image statistics with higher-level properties, visual realism of composite images is further considered in [20, 47].

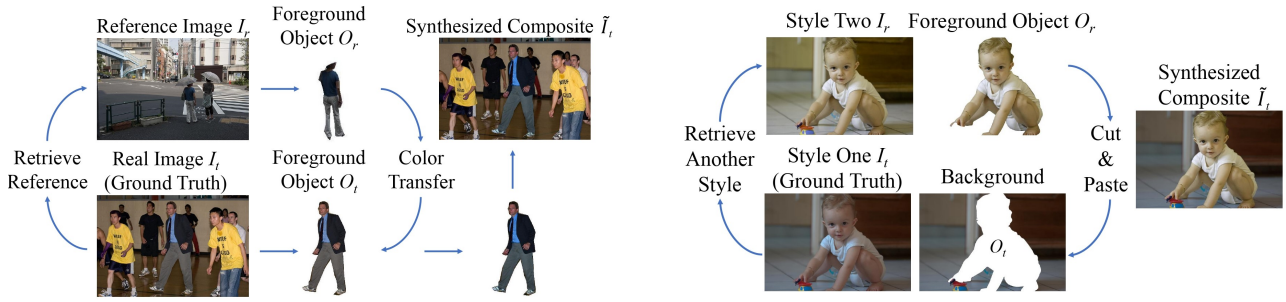
Recently, Zhu *et al.* [54] trained a CNN model to perform realism assessment of composite images and applied the model to improve realism. Tsai *et al.* [43] proposed the first end-to-end CNN network to directly produce harmonized images, in which an extra segmentation branch is used to incorporate semantic information. In [45], an attention module was proposed to learn the attended foreground and background features separately. Different from these existing methods, our proposed method aims to translate the foreground domain to the background domain by using a domain verification discriminator.

Image-to-Image Translation: A variety of tasks that map an input image to a corresponding output image are collectively named image-to-image translation, such as image super-resolution [15, 16, 22], inpainting [33, 50], colorization [51, 21], denoising [26], de-blurring [46], dehazing [38, 3], demo-saicking [8], decompression [6], and few-shot image generation [10]. However, there are still limited deep-learning based research in image harmonization field.

Moreover, several general frameworks of image-to-image translation have also been proposed [11, 27, 49]. For the tasks with paired training data, Among them, paired GANs like [11] designed for paired training data can be applied to image harmonization, but they do not consider the uniqueness of image harmonization problem. Our model extends paired GAN with a domain verification discriminator, which goes beyond conventional paired GAN.

3. Dataset Construction

In this section, we will fully describe the data acquisition process to build our dataset iHarmony4. Based on real images, we first generate composite images and then filter out the unqualified composite images.



(a) Microsoft COCO & Flickr (b) MIT-Adobe Fivek & day2night

Figure 1: The illustration of our data acquisition process. (a) On Miscrosoft COCO and Flickr datasets, given a target image I_t with foreground object O_t, we find a reference image I_r with foreground object O_r from the same category as O_t, and then transfer color information from O_r to O_t. (b) On MIT-Adobe5k and day2night datasets, given a target image I_t with foreground object O_t, we find its another version I_r (edited to present a different style or captured in a different condition) and overlay O_t with the corresponding O_r at the same location in I_r.

3.1. Composite Image Generation

The process of generating synthesized composite image from a real image has two steps: foreground segmentation and foreground adjustment, as illustrated in Figure 1.

Foreground Segmentation: For COCO dataset, we use the provided segmentation masks for 80 categories. The other datasets (*i.e.*, Adobe5k, Flickr, and day2night) are not associated with segmentation masks, so we manually segment one or more foreground regions for each image.

On all four sub-datasets, we ensure that each foreground region occupies a reasonable area of the whole image and also attempt to make the foreground objects cover a wide range of categories.

Foreground Adjustment: After segmenting a foreground region O_t in one image I_t, we need to adjust the appearance of O_t. For ease of description, I_t is dubbed as target image. As suggested in [43], another image I_r containing the foreground region O_r is chosen as reference image. Then, color information is transferred from O_r to O_t, leading to a synthesized composite image $\tilde{I}_t$.

For Adobe5k dataset, each real image is retouched by five professional photographers, so one real target image I_t is accompanied by five edited images $\{I_i|_{i=1}^5\}$ in different styles. We could randomly select I_r from $\{I_i|_{i=1}^5\}$ and overlay O_t in I_t with the corresponding region O_r at the same location in I_r.

For day2night dataset, each scene is captured in different conditions, resulting in a series of aligned images $\{I_i|_{i=1}^n\}$. Similar to Adobe5k, a target image I_t and a reference image I_r could be randomly selected from $\{I_i|_{i=1}^n\}$, followed by overlaying O_t in I_t with the corresponding region O_r in I_r. However, different from Adobe5k, we need to make sure that O_t and O_r are the same object without essential change. For example, moving objects (*e.g.*, person, animal, car) in I_t may move or disappear in I_r. Besides, even the static objects (*e.g.* building, mountain) in I_t may be different from those in I_r, like building with lights on in I_t while lights off in I_r. The above foreground changes come from the objects themselves instead of the capture condition, and thus we exclude those pairs from our dataset.

For COCO and Flickr datasets, since they do not have aligned images, given a target image I_t with foreground O_t, we randomly select a reference image I_r with foreground O_r belonging to the same category as O_t. For COCO dataset with segmentation annotations for 80 categories, given I_t in COCO, we retrieve I_r from COCO itself. For Flickr dataset without segmentation annotations, we use ADE20K pretrained scene-parsing model [52] to obtain the dominant category label of O_t and retrieve I_r from ADE20K dataset [52]. Then, as suggested in [43], we apply color transfer method to transfer color information from O_r to O_t. Nevertheless, the work [43] only utilizes one color transfer method [23], which limits the diversity of generated images. Considering that color transfer methods can be categorized into four groups based on parametric/non-parametric and correlated/decorrelated color space, we select one representative method from each group, *i.e.*, parametric method [37] (*resp.*, [44]) in decorrelated (*resp.*, correlated) color space and non-parametric method [7] (*resp.*, [36]) in decorrelated (*resp.*, correlated) color space. Given a pair of O_t and O_r, we randomly choose one from the above four color transfer methods.

3.2. Composite Image Filtering

Through foreground segmentation and adjustment, we can obtain a large amount of synthesized composite images. However, some of the synthesized foreground objects look unrealistic, so we use aesthetics prediction model [17] to remove unrealistic composite images. To further remove

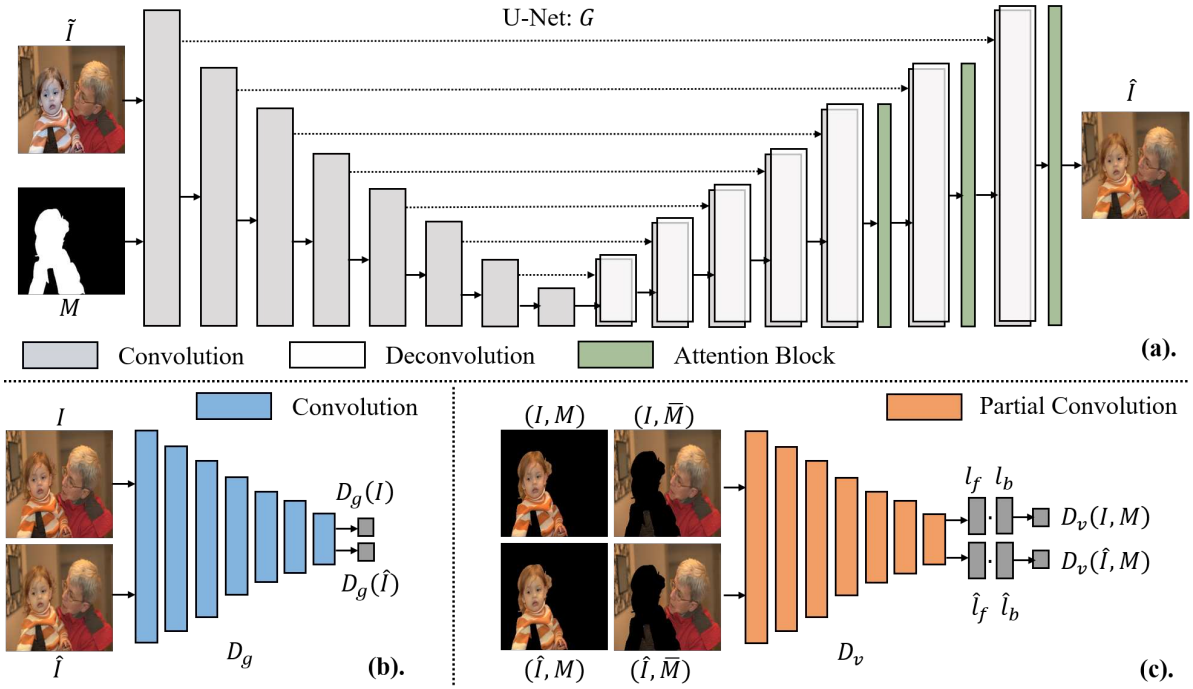


Figure 2: Illustration of DoveNet architecture, which consists of (a) attention enhanced U-Net generator, (b) global discriminator, and (c) our proposed domain verification discriminator.

unrealistic composite images, we train a binary CNN classifier by using the real images as positive samples and the unrealistic composite images identified by [17] as negative samples. When training the classifier, we also feed foreground masks into CNN for better performance.

After two steps of automatic filtering, there are still some remaining unrealistic images. Thus, we ask human annotators to remove the remaining unrealistic images manually. During manual filtering, we also consider another two critical issues: 1) for COCO dataset, some selected foreground regions are not very reasonable such as highly occluded objects, so we remove these images; 2) for COCO and Flickr datasets, the hue of some foreground objects are vastly changed after color transfer, which generally happens to the categories with large intra-class variance. For example, a red car is transformed into a blue car, or a man in red T-shirt is transformed into a man in green T-shirt. This type of color transfer is not very meaningful for image harmonization task, so we also remove these images.

3.3. Differences between Our Dataset and [43]

Our dataset iHarmony4 is an augmented and enhanced version of the dataset in [43]: 1) Our dataset contains an additional sub-dataset Hday2night, which is not considered in [43]. Unlike the other three sub-datasets, Hday2night consists of real composite images, which is closer to real-world application; 2) Besides, we also attempt to address some issues not considered in [43], such as the diversity and quality issues of synthesized composite images; 3) We apply both well-designed automatic filtering and deliberate manual filtering to guarantee the high quality of our dataset.

4. Our Method

Given a real image I, we have a corresponding composite image $\tilde{I}$, where the foreground mask M indicates the region to be harmonized and the background mask is $\bar{M} = 1 - M$. Our goal is to train a model that reconstructs I with a harmonized image $\hat{I}$, which is expected to be as close to I as possible.

We leverage the GAN [9] framework to generate plausible and harmonious images. As demonstrated in Figure 2, in DoveNet, we use an attention enhanced U-Net generator G, which takes $(\tilde{I}, M)$ as inputs and outputs a harmonized image $\hat{I}$. Besides, we use two different discriminators D_g and D_v to guide G for generating more realistic and harmonious images. The first discriminator D_g is a traditional global discriminator, which discriminates real images and generated images. The second discriminator D_v is our proposed domain verification discriminator, which verifies whether the foreground and background of a given image come from the same domain.

4.1. Attention Enhanced Generator

Our generator G is based on U-Net [39] with skip links from encoders to decoders. Inspired by [45], we leverage attention blocks to enhance U-Net. Specifically, we first concatenate encoder and decoder features, based on which full attention [48] (integration of spatial attention and channel attention) is learnt for encoder feature and decoder feature separately. Then, we concatenate the attended encoder and decoder features. In total, we insert three attention blocks into U-Net as depicted in Figure 2 and the details of attention block can be found in Supplementary. We enforce the generated image $\hat{I} = G(\tilde{I}, M)$ to be close to ground-truth real image I by $L_{rec} = \|\hat{I} - I\|_1$.

4.2. Global Discriminator

The global discriminator D_g is designed to help G generate plausible images, which takes I as real images and $\hat{I}$ as fake images. Following [28], we apply spectral normalization after each convolutional layer and leverage hinge loss for stabilizing training, which is given by

$$\begin{aligned} L_{D_g} &= \mathbb{E}[\max(0, 1 - D_g(I))] + \mathbb{E}[\max(0, 1 + D_g(\hat{I}))], \\ L_{G_g} &= -\mathbb{E}[D_g(G(\tilde{I}, M))]. \end{aligned} \tag{1}$$

When training D_g by minimizing L_{D_g}, D_g is encouraged to produce large (*resp.*, small) scores for real (*resp.*, generated) images. While training G by minimizing L_{G_g}, the generated images are expected to fool D_g and obtain large scores.

4.3. Domain Verification Discriminator

Besides the global discriminator, we also design a domain verification discriminator to verify whether the foreground and background of a given image belong to the same domain. As discussed in Section 1, the foreground and background of a real (*resp.*, composite) image are captured in the same condition (*resp.*, different conditions), and thus belong to the same domain (*resp.*, different domains), which is dubbed as a positive (*resp.*, negative) foreground-background pair.

To extract domain representation for foreground and background, we adopt partial convolution [25], which is well-tailored for image harmonization task. Partial convolution only aggregates the features from masked regions, which can avoid information leakage from unmasked regions or invalid information corruption like zero padding. Our domain representation extractor F is formed by stacking partial convolutional layers, which leverages the advantage of partial convolution to extract domain information for foreground and background separately.

Formally, given a real image I, let $I_f = I \circ M$ (*resp.*, $I_b = I \circ \bar{M}$) be the masked foreground (*resp.*, background) image, in which $\circ$ means element-wise product. Domain representation extractor $F(I_f, M)$ (*resp.*, $F(I_b, \bar{M})$) extracts the foreground representation l_f (*resp.*, l_b) based on I_f (*resp.*, I_b) and M (*resp.*, $\bar{M}$). Similarly, given a harmonized image $\hat{I}$, we apply the same domain representation extractor F to extract its foreground representation $\hat{l}_f$ and background representation $\hat{l}_b$.

After obtaining domain representations, we calculate the domain similarity $D_v(I, M) = l_f \cdot l_b$ (*resp.*, $D_v(\hat{I}, M) = \hat{l}_f \cdot \hat{l}_b$) as the verification score for the real (*resp.*, generated) images, where $\cdot$ means inner product. In analogy to (1), the loss functions *w.r.t.* the domain verification discriminator can be written as

$$\begin{aligned} L_{D_v} &= \mathbb{E}[\max(0, 1 - D_v(I, M))] \\ &\quad + \mathbb{E}[\max(0, 1 + D_v(\hat{I}, M))], \\ L_{G_v} &= -\mathbb{E}[D_v(G(\tilde{I}, M), M)]. \end{aligned} \tag{2}$$

When training D_v by minimizing L_{D_v}, D_v is encouraged to produce large (*resp.*, small) scores for positive (*resp.*, negative) foreground-background pairs. While training G by minimizing L_{G_v}, the generated images are expected to fool D_v and obtain large scores. By matching the foreground domain with the background domain, the generated images are expected to have compatible foreground and background. So far, the total loss function for training generator G is

$$L_G = L_{rec} + \lambda(L_{G_g} + L_{G_v}), \tag{3}$$

in which the trade-off parameter λ is set as 0.01 in our experiments. Similar to GAN [9], we update generator G and two discriminators D_g, D_v alternatingly. Due to the usage of DOmain VErification (DOVE) discriminator, we name our method as DoveNet.

5. Experiments

In this section, we analyze the statistics of our constructed iHarmony4 dataset. Then, we evaluate baselines and our proposed DoveNet on our constructed dataset.

5.1. Dataset Statistics

HCOCO: Microsoft COCO dataset [24] contains 118k images for training and 41k for testing. It provides the object segmentation masks for each image with 80 object categories annotated in total. To generate more convincing composites, training set and test set are merged together to guarantee a wider range of available references. Based on COCO dataset, we build our HCOCO sub-dataset with 42828 pairs of synthesized composite image and real image.
HAdobe5k: MIT-Adobe5k dataset [2] covers a wide range of scenes, objects, and lighting conditions. For all the 5000 photos, each of them is retouched by five photographers,

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	172.47	31.63
Lalonde and Efros[20]	110.10	31.14	158.90	29.66	329.87	26.43	199.93	29.80	150.53	30.16
Xue *et al.*[47]	77.04	33.32	274.15	28.79	249.54	28.32	190.51	31.24	155.87	31.40
Zhu *et al.*[54]	79.82	33.04	414.31	27.26	315.42	27.52	136.71	32.32	204.77	30.72
DIH [43]	51.85	34.69	92.65	32.28	163.38	29.55	82.34	34.62	76.77	33.41
S^2AM [45]	41.07	35.47	63.40	33.77	143.45	30.03	76.61	34.50	59.67	34.35
DoveNet	**36.72**	**35.83**	**52.32**	**34.34**	**133.14**	**30.21**	**54.05**	**35.18**	**52.36**	**34.75**

Table 1: Results of different methods on our four sub-datasets. The best results are denoted in boldface.

Sub-dataset	HCOCO	HAdobe5k	HFlickr	Hday2night
#Training	38545	19437	7449	311
#Test	4283	2160	828	133

Table 2: The numbers of training and test images on our four sub-datasets.

producing five different renditions. We use 4329 images with one segmented foreground object in each image to build our HAdobe5k sub-dataset, resulting in 21597 pairs of synthesized composite image and real image.

HFlickr: Flickr website is a public platform for uploading images by amateur photographers. We construct our HFlickr sub-dataset based on crawled 4833 Flickr images with one or two segmented foreground object in each image. Our HFlickr sub-dataset contains 8277 pairs of synthesized composite image and real image.

Hday2night: Day2night dataset [53] collected from AMOS dataset [13] contains images taken at different times of the day with fixed webcams. There are 8571 images of 101 different scenes in total. We select 106 target images from 80 scenes with one segmented foreground object in each image to generate composites. Due to the stringent requirement mentioned in Section 3.1, we only obtain 444 pairs of synthesized composite image and real image, without degrading the dataset quality.

For each sub-dataset (*i.e.*, HCOCO, HAdobe5k, HFlickr, and Hday2night), all pairs are split into training set and test set. We ensure that the same target image does not appear in the training set and test set simultaneously, to avoid that the trained model simply memorize the target image. The numbers of training and test images in four sub-datasets are summarized in Table 2. The sample images and other statistics are left to Supplementary due to space limitation.

5.2. Implementation Details

Following the network architecture in [12], we apply eight downsample blocks inside the generator, in which each block contains a convolution with a kernel size of four and stride of two. After the convolution layers, we apply LeakyReLU activation and instance normalization layer. We use eight deconvolution layers to upsample the feature to generate images. For global (*resp.*, verification) discriminator, we use seven convolutional (*resp.*, partial convolutional) layers and LeakyReLU is applied after all the convolutional layers before the last one in both discriminators. We use Adam optimizer with learning rate 0.002. Following [43], we use Mean-Squared Errors (MSE) and PSNR scores on RGB channels as the evaluation metric. We report the average of MSE and PSNR over the test set. We resize the input images as 256×256 during both training and testing. MSE and PSNR are also calculated based on 256×256 images.

5.3. Comparison with Existing Methods

We compare with both traditional methods [20, 47] and deep learning based methods [54, 43, 45]. Although Zhu *et al.* [54] is a deep learning based method, it relies on the pretrained aesthetic model and does not require our training set. DIH [43] originally requires training images with segmentation masks, which are not available in our problem. Therefore, we compare with DIH by removing its semantic segmentation branch, because we focus on pure image harmonization task without using any auxiliary information. For all baselines, we conduct experiments with their released code if available, and otherwise based on our own implementation.

Following [43], we merge the training sets of all four sub-datasets as a whole training set to learn the model, which is evaluated on the test set of each sub-dataset and the whole test set. The results of different methods are summarized in Table 1, from which we can observe that deep learning based methods using our training set [43, 45] are generally better than traditional methods [20, 47], which demonstrates the effectiveness of learning to harmonize images from paired training data. We also observe that S^2AM is better than DIH, which shows the benefit of its proposed attention block. Our DoveNet outperforms all the baselines by a large margin and achieves the best results on all four sub-datasets, which indicates the advantage of our domain verification discriminator.

Sub-dataset	HCOCO		HAdobe5k		HFlickr		Hday2night		All	
Evaluation metric	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑
U-Net	46.87	34.30	77.16	32.34	160.17	29.25	57.60	34.25	68.57	33.16
U-Net+att	43.13	35.15	57.52	33.83	159.99	29.56	56.40	34.89	61.15	34.13
U-Net+att+adv	38.44	35.54	54.56	34.08	143.03	29.99	55.68	34.72	55.15	34.48
U-Net+att+ver	39.79	35.33	53.84	34.19	136.60	30.04	55.64	34.94	55.00	34.40
U-Net+att+adv+ver	**36.72**	**35.83**	**52.32**	**34.34**	**133.14**	**30.21**	**54.05**	**35.18**	**52.36**	**34.75**

Table 3: Results of our special cases on our four sub-datasets. U-Net is the backbone generator. "att" stands for our used attention block, "adv" stands for the adversarial loss of global discriminator. "ver" stands for the verification loss of our proposed verification discriminator. The best results are denoted in boldface.

Foreground ratios	$0\% \sim 5\%$		$5\% \sim 15\%$		$15\% \sim 100\%$		$0\% \sim 100\%$	
Evaluation metric	MSE↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓
Input composite	28.51	1208.86	119.19	1323.23	577.58	1887.05	172.47	1387.30
Lalonde and Efros[20]	41.52	1481.59	120.62	1309.79	444.65	1467.98	150.53	1433.21
Xue *et al.*[47]	31.24	1325.96	132.12	1459.28	479.53	1555.69	155.87	1411.40
Zhu *et al.*[54]	33.30	1297.65	145.14	1577.70	682.69	2251.76	204.77	1580.17
DIH [43]	18.92	799.17	64.23	725.86	228.86	768.89	76.77	773.18
S^2AM [45]	15.09	623.11	48.33	540.54	177.62	592.83	59.67	594.67
DoveNet	**14.03**	**591.88**	**44.90**	**504.42**	**152.07**	**505.82**	**52.36**	**549.96**

Table 4: 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.

Method	B-T score↑
Input composite	0.624
Lalonde and Efros [20]	0.260
Xue *et al.* [47]	0.567
Zhu *et al.* [54]	0.337
DIH [43]	0.948
S^2AM [45]	1.229
DoveNet	**1.437**

Table 5: B-T scores of different methods on 99 real composite images provided in [43].

5.4. Ablation Studies

In this section, we first investigate the effectiveness of each component in our DoveNet, and then study the impact of foreground ratio on the harmonization performance.

First, the results of ablating each component are reported in Table 3. By comparing "U-Net" with DIH in Table 1, we find that our backbone generator is better than that used in DIH [43]. We also observe that "U-Net+att" outperforms "U-Net", which shows the benefit of using attention block. Another observation is that "U-Net+att+adv" (*resp.*, "U-Net+att+ver") performs more favorably than "U-Net+att", which indicates the advantage of employing global discriminator (*resp.*, our domain verification discriminator). Finally, our full method, *i.e.*, "U-Net+att+adv+ver", achieves the best results on all four sub-datasets.

Second, our dataset has a wide range of foreground ratios (the area of foreground over the area of whole image) in which the foreground ratios of most images are in the range of $[1\%, 90\%]$ (see Supplementary). Here, we study the impact of different foreground ratios on the harmonization performance. Especially when the foreground ratio is very small, the reconstruction error of background may overwhelm the harmonization error of foreground. Therefore, besides MSE on the whole image, we introduce another evaluation metric: foreground MSE (fMSE), which only calculates the MSE in the foreground region. We divide foreground ratios into three ranges, *i.e.*, $0\% \sim 5\%$, $5\% \sim 15\%$, and $15\% \sim 100\%$. We adopt such a partition because more images have relatively small foreground ratios. Then, we report MSE and fMSE of different methods for each range on the whole test set in Table 4. Obviously, MSE increases as the foreground ratio increases. Based on Table 4, DoveNet outperforms all the baselines *w.r.t.* MSE and fMSE in each range of foreground ratios, especially when the foreground ratio is large, which demonstrates the robustness of our method.

5.5. Qualitative Analyses

In Figure 3, we show the ground-truth real image, input composite image, as well as the harmonized images generated by DIH [43], S^2AM[45], DoveNet (w/o ver),

Figure 3: Example results of different methods on our four sub-datasets. From top to bottom, we show one example from our HAdobe5k, HCOCO, Hday2night, and HFlickr sub-dataset respectively. From left to right, we show the ground-truth real image, input composite image, DIH [43], S^2AM[45], our special case DoveNet (w/o ver) and our full method DoveNet.

and DoveNet. DoveNet (w/o ver) corresponds to "U-Net+att+adv" in Table 3, which removes domain verification discriminator from our method. We observe that our proposed method could produce the harmonized images which are more harmonious and closer to the ground-truth real images. By comparing DoveNet (w/o ver) and DoveNet, it can be seen that our proposed verification discriminator is able to push the foreground domain close to the background domain, leading to better-harmonized images.

5.6. User Study on Real Composite Images

We further compare our proposed DoveNet with baselines on 99 real composited images used in [43]. Because the provided 99 real composited images do not have ground-truth images, it is impossible to compare different methods quantitatively using MSE and PSNR. Following the same procedure in [43], we conduct user study on the 99 real composited images for subjective evaluation. Specifically, for each real composite image, we can obtain 7 outputs, including the original composite image and the harmonized images of 6 methods (see Table 1). For each real composite image, we can construct pairs of outputs by selecting from 7 outputs. Then, we invite 50 human raters to see a pair of outputs at a time and ask him/her to choose the more realistic and harmonious one. A total of 51975 pairwise results are collected for all 99 real composite images, in which 25 results are obtained for each pair of outputs on average. Finally, we use the Bradley-Terry model (B-T model) [1, 19] to calculate the global ranking score for each method and report the results in Table 5.

From Table 5, we have similar observation as in Table 1. In particular, deep learning based methods using our training set are generally better than traditional methods, among which DoveNet achieves the highest B-T score. To visualize the comparison, we put the results of different methods on all 99 real composite images in Supplementary.

6. Conclusions

In this work, we have contributed an image harmonization dataset iHarmony4 with four sub-datasets: HCOCO, HAdobe5k, HFlickr, and Hday2night. We have also proposed DoveNet, a novel deep image harmonization method with domain verification discriminator. Extensive experiments on our dataset have demonstrated the effectiveness of our proposed method.

Acknowledgement

The work is supported by the National Key R&D Program of China (2018AAA0100704) and is partially sponsored by National Natural Science Foundation of China (Grant No.61902247) and Shanghai Sailing Program (19YF1424400).

References

[1] 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. 8

[2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR*, 2011. 1, 5

[3] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. 2

[4] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. Color harmonization. 25(3):624–630, 2006. 2

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 1

[6] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *ICCV*, 2015. 2

[7] Ulrich Fecker, Marcus Barkowsky, and André Kaup. Histogram-based prefiltering for luminance and chrominance compensation of multiview video. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(9):1258–1267, 2008. 3

[8] Michaël Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand. Deep joint demosaicking and denoising. *ACM Transactions on Graphics*, 35(6):191, 2016. 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2, 4, 5

[10] Yan Hong, Li Niu, Jianfu Zhang, and Liqing Zhang. MatchingGAN: Matching-based few-shot image generation. In *ICME*, 2020. 2

[11] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 6

[13] N. Jacobs, N. Roman, and R. Pless. Consistent temporal variations in many outdoor scenes. In *CVPR*, 2007. 6

[14] Jiaya Jia, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. Drag-and-drop pasting. *ACM Transactions on Graphics*, 25(3):631–637, 2006. 2

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 2

[16] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016. 2

[17] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016. 3, 4

[18] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics*, 33(4), 2014. 2

[19] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *CVPR*, 2016. 8

[20] Jean-Francois Lalonde and Alexei A Efros. Using color compatibility for assessing image realism. In *ICCV*, 2007. 1, 2, 6, 7

[21] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *ECCV*, 2016. 2

[22] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2

[23] Joon-Young Lee, Kalyan Sunkavalli, Zhe Lin, Xiaohui Shen, and In So Kweon. Automatic content-aware color and tone stylization. In *CVPR*, 2016. 3

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 5

[25] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018. 5

[26] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *NeurIPS*, 2016. 2

[27] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[28] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 5

[29] Li Niu, Jianfei Cai, and Dong Xu. Domain adaptive fisher vector for visual recognition. In *ECCV*, 2016. 2

[30] Li Niu, Wen Li, and Dong Xu. Multi-view domain generalization for visual recognition. In *ICCV*, 2015. 2

[31] Li Niu, Wen Li, and Dong Xu. Visual recognition by learning from web data: A weakly supervised domain generalization approach. In *CVPR*, 2015. 2

[32] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015. 2

[33] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2

[34] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Transactions on graphics*, 22(3):313–318, 2003. 2

[35] F. Pitie, A. C. Kokaram, and R. Dahyot. N-dimensional probability density function transfer and its application to color transfer. In *ICCV*, 2005. 2

[36] François Pitié, Anil C Kokaram, and Rozenn Dahyot. Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 107(1-2):123–137, 2007. 3

[37] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001. 2, 3

[38] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *ECCV*, 2016. 2

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *ICM*, 2015. 5

[40] Yichang Shih, Sylvain Paris, Frédo Durand, and William T Freeman. Data-driven hallucination of different times of day from a single outdoor photo. *ACM Transactions on Graphics*, 32(6):200, 2013. 2

[41] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM Transactions on Graphics*, 29(4):125, 2010. 2

[42] Michael W Tao, Micah K Johnson, and Sylvain Paris. Error-tolerant image compositing. In *ECCV*, 2010. 2

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

[44] Xuezhong Xiao and Lizhuang Ma. Color transfer in correlated color space. In *VRCAI*, 2006. 3

[45] Cun Xiaodong and Pun Chi-Man. Improving the harmony of the composite image by spatial-separated attention module. *arXiv preprint arXiv:1907.06406*, 2019. 1, 2, 5, 6, 7, 8

[46] Li Xu, Jimmy S. J. Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. In *NeurIPS*, 2014. 2

[47] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. Understanding and improving the realism of image composites. *ACM Transactions on Graphics*, 31(4):84, 2012. 1, 2, 6, 7

[48] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 5

[49] Jianfu Zhang, Yuanyuan Huang, Yaoyi Li, Weijie Zhao, and Liqing Zhang. Multi-attribute transfer via disentangled representation. In *AAAI*, 2019. 2

[50] Jianfu Zhang, Li Niu, Dexin Yang, Liwei Kang, Yaoyi Li, Weijie Zhao, and Liqing Zhang. GAIN: gradient augmented inpainting network for irregular holes. In *MM*, 2019. 2

[51] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2

[52] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 3

[53] Hao Zhou, Torsten Sattler, and David W Jacobs. Evaluating local features for day-night matching. In *ECCV*, 2016. 2, 6

[54] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A Efros. Learning a discriminative model for the perception of realism in composite images. In *ICCV*, 2015. 1, 2, 6, 7