

Mask-ShadowGAN: Learning to Remove Shadows from Unpaired Data

Xiaowei Hu¹, Yitong Jiang², Chi-Wing Fu^{1,2,*}, and Pheng-Ann Heng^{1,2,*}

¹ Department of Computer Science and Engineering, The Chinese University of Hong Kong

² Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

Abstract

This paper presents a new method for shadow removal using unpaired data, enabling us to avoid tedious annotations and obtain more diverse training samples. However, directly employing adversarial learning and cycle-consistency constraints is insufficient to learn the underlying relationship between the shadow and shadow-free domains, since the mapping between shadow and shadow-free images is not simply one-to-one. To address the problem, we formulate Mask-ShadowGAN, a new deep framework that automatically learns to produce a shadow mask from the input shadow image and then takes the mask to guide the shadow generation via re-formulated cycle-consistency constraints. Particularly, the framework simultaneously learns to produce shadow masks and learns to remove shadows, to maximize the overall performance. Also, we prepared an unpaired dataset for shadow removal and demonstrated the effectiveness of Mask-ShadowGAN on various experiments, even it was trained on unpaired data.

1. Introduction

Shadow removal is a very challenging task. We have to remove the shadows, and simultaneously, restore the background behind the shadows. Particularly, shadows have a wide variety of shapes over a wide variety of backgrounds. Recently, learning-based methods [10, 11, 18, 32], especially those using deep learning [13, 27, 35], have become the de facto standard for shadow removal, given their remarkable performance. These methods are typically trained on pairs of shadow and shadow-free images in a supervised manner, where the paired data is prepared by taking a photo with shadows and then taking another photo of the scene without shadows by removing the associated objects.

Such approach to prepare training data has several limitations. First, it is very tedious to prepare the training data, since for each scene, we need to manually fix the camera

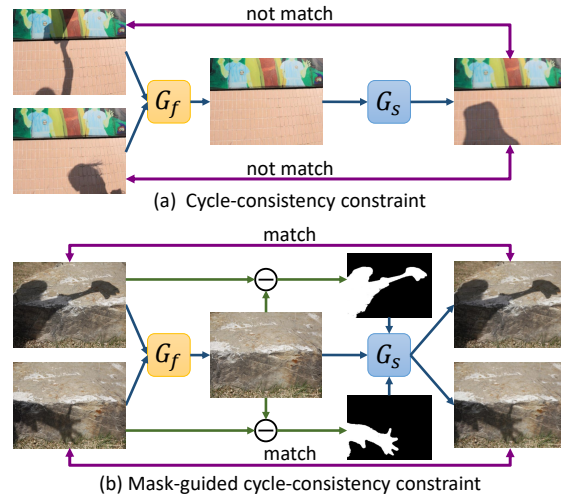


Figure 1: Directly using (a) the cycle-consistency constraint is insufficient; the generator G_s cannot produce different shadow image outputs for different inputs. (b) Mask-ShadowGAN learns a shadow mask from the input to guide G_s to generate shadow images that better match the inputs.

and then add & remove objects to obtain a pair of shadow and shadow-free images. Moreover, the approach *limits the kinds of scenes that data can be prepared*, since it is hard to capture shadow-free images for shadows casted by large objects such as trees and buildings. Lastly, *the training pairs may have inconsistent colors and luminosity, or shift in camera views*, since the camera exposure and pose, as well as the environmental lighting, may vary when we take the photo pair with and without the shadows.

To address these problems, we present a new approach to learn to remove shadows from *unpaired training data*. Our key idea is to learn the underlying relationship between a shadow domain \mathbb{D}_s (a set of real images with shadows) and a shadow-free domain \mathbb{D}_f (a set of real shadow-free images), where we do not have any explicit association between individual images in \mathbb{D}_s and \mathbb{D}_f . Here, we want to train a network G_f , which takes a shadow image as input and produces an output image that is indistinguishable from the shadow-free images in \mathbb{D}_f , by adversarial learn-

*Co-corresponding authors

ing [9, 15]. This mapping is highly under-constrained, so the network can easily collapse during the training [43]. Hence, we train another network G_s to learn the inverse mapping, *i.e.*, G_s , to translate a shadow-free image into a shadow image like those in \mathbb{D}_s , and impose the following cycle-consistency constraints [43] on images $I_s \in \mathbb{D}_s$ and $I_f \in \mathbb{D}_f$, *i.e.*, $G_s(G_f(I_s))$ should be the same as the input shadow image I_s , and $G_f(G_s(I_f))$ should be the same as the input shadow-free image I_f .

Fundamentally, deep neural networks generate a unique output for the same input. Having said that, for the same shadow-free image, G_s always generates the same shadow image with the same shadow shape. However, this is clearly insufficient to shadow generation, since for the same background, we may have different shadows. Figure 1(a) shows a further illustration: starting from different shadow images with the same background, while G_f produces the same shadow-free image, G_s will fail to generate different outputs that match the corresponding original inputs. Hence, the cycle-consistency constraint cannot hold, and we cannot train G_f and G_s to learn to remove and generate shadows.

To address the problem, we formulate a mask-guided generative adversarial network, namely *Mask-ShadowGAN*, which learns to produce a shadow mask from an input shadow image during the training and takes the mask to guide G_s to generate shadows. Therefore, given a shadow-free image, we can generate different shadows by using different shadow masks. Further, from an input shadow image, we can produce a shadow-free image and generate suitable shadows on the image to produce an output that matches the corresponding input; see Figure 1(b). Hence, we can adopt the cycle-consistency constraint to train G_f and G_s .

To our best knowledge, this is the first data-driven method, which trains a network with unpaired data, for shadow removal. Particularly, we design *Mask-ShadowGAN* to learn to remove shadows from unpaired data, bringing forth the advantage of using more shadow and shadow-free images for network learning. Also, we prepare the first unpaired shadow-and-shadow-free image dataset with diverse scenes. Last, we perform various experiments to evaluate *Mask-ShadowGAN* and demonstrate its effectiveness, even it was trained on unpaired data. Results show that our method produces comparable performance with existing works on the existing benchmarks, and outperforms others for more general shadow images without paired ground truth images. The source code and dataset are publicly available at <https://xw-hu.github.io/>.

2. Related Work

2.1. Shadow Removal

Early methods remove shadows by modeling images as combinations of shadow and shadow-free layers [2, 4, 5, 6,

7, 23, 24, 40], or by transferring colors from non-shadow to shadow regions [28, 37, 38, 39]. Since the underlying shadow models are not physically correct, they usually cannot handle shadows in complex real scenes [18]. Later, statistical-learning methods were explored to find and remove shadows using features, such as intensity [8, 10, 11], color [11, 32], texture [11, 32], and gradient [10]. However, these hand-crafted features lack high-level semantics for understanding the shadows. Later, Khan *et al.* [17, 18] adopted convolutional neural networks (CNNs) to detect shadows followed by a Bayesian model to remove shadows.

In recent years, shadow removal mainly relies on CNNs that are trained end-to-end to learn the mapping from paired shadow and shadow-free images. Qu *et al.* [27] developed three subnetworks to extract features from multiple views and embedded these subnetworks in a framework to remove shadows. Wang *et al.* [35] used a conditional generative adversarial network (CGAN) to detect shadows and another CGAN to remove shadows. Hu *et al.* [13, 14] explored the direction-aware spatial context to detect and remove shadows. However, these methods are all trained on paired images, which incur several limitations, as discussed in the introduction. Similarly, the recent shadow detection methods [33, 21, 44, 42] also trained their deep neural networks using paired data. Unlike previous works, we present a new framework based on adversarial learning to learn to remove shadows from unpaired training data.

2.2. Unsupervised Learning

Unsupervised learning receives great attention in recent years. These methods can roughly be divided into three approaches. The first approach learns the feature representations by generating images, where early methods, such as the auto-encoder [12] and denoising auto-encoder [34], encoded the input images in a latent space by reconstructing them with a low error. Some more recent works transferred the synthetic images into “real” images through adversarial learning [29]. The second approach is self-supervised learning, which learns the invariant features by designing the auxiliary training objectives via labels that are free to obtain, *e.g.*, Doersch *et al.* [3] predicted the location of image patches for feature learning; Pathak *et al.* [26] learned the feature representations from segmented moving objects in videos; and Wang *et al.* [36] learned the visual representations from the transitive relations between images.

The last approach is more related to our work. It learns the underlying mapping between domains in the form of unpaired data. CycleGAN [43] and other similar models [19, 41] used two generative adversarial networks [9] to formulate the cycle consistency constraints, and learned the mapping to translate images between domains. Though they can learn an arbitrary one-to-one mapping, a trained network can only produce a single output for the same in-

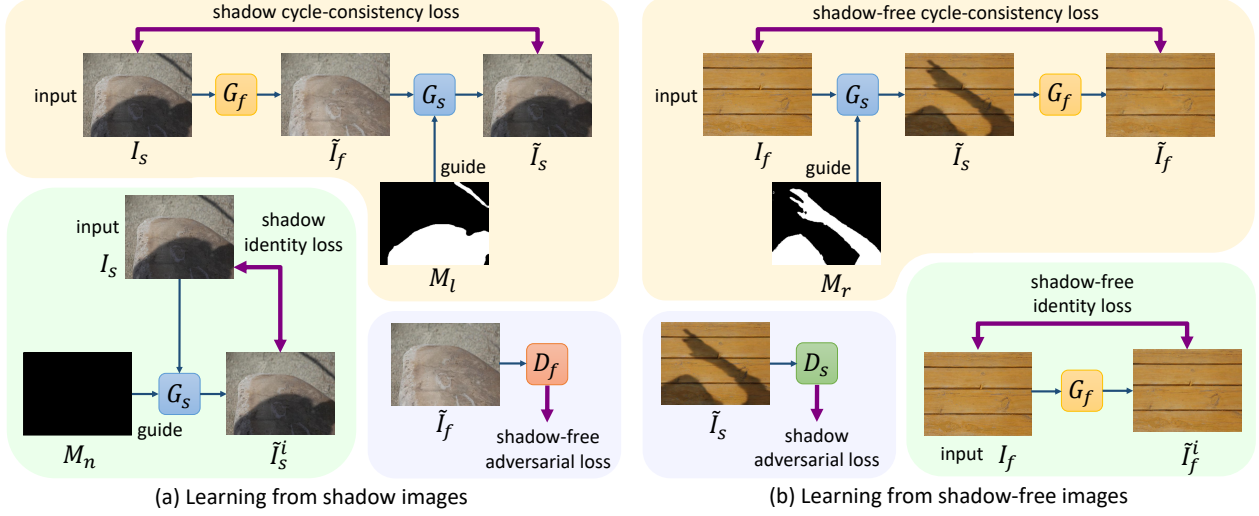


Figure 2: The schematic illustration of our Mask-ShadowGAN, which has two parts: (a) one to learn from real shadow images and (b) the other to learn from real shadow-free images. Each part includes three losses: cycle-consistency loss (yellow), identity loss (green), and adversarial loss (blue). Besides, G_f and G_s denote the generators, which produce the shadow-free and shadow images while D_f and D_s are the discriminators to determine the whether the generated images are real shadow-free or shadow images. I_s and I_f are the real shadow and shadow-free image; \tilde{I}_s and \tilde{I}_s^i denote the generated shadow images while \tilde{I}_f and \tilde{I}_f^i denote the generated shadow-free images; M_n , M_l and M_r are the shadow masks.

put image. To extend cycle consistency to handle many-to-many mapping, Almahairi *et al.* [1] and Lee *et al.* [22] explored the use of latent variables to generate diverse outputs. In contrast, the relationship between shadow and shadow-free images can be explicitly modelled by their difference. Hence, in this work, we design a framework to learn to produce the shadow mask for guiding the shadow generation, and learn to use the difference between shadow and shadow-free images in the shadow regions to model the relationship between the shadow and shadow-free images.

3. Methodology

Figure 2 outlines the overall network architecture of our Mask-ShadowGAN framework, which has two parts: one to learn from real shadow images (Section 3.1) and the other to learn from real shadow-free images (Section 3.2).

3.1. Learning from Shadow Images

Starting from a real shadow image I_s , we first use a generator network G_f to transform it into a shadow-free image \tilde{I}_f . Then, we use an adversarial discriminator D_f to differentiate whether \tilde{I}_f is a real shadow-free image or not:

$$\tilde{I}_f = G_f(I_s), D_f(\tilde{I}_f) = \text{real or fake?} \quad (1)$$

Here, we optimize the following objective function, simultaneously for the generator and its discriminator:

$$L_{GAN}^a(G_f, D_f) = \mathcal{E}_{I_f \sim p_{data}(I_f)}[\log(D_f(I_f))] + \mathcal{E}_{I_s \sim p_{data}(I_s)}[\log(1 - D_f(G_f(I_s)))] \quad (2)$$

where \mathcal{E} denotes the error; p_{data} denotes the data distribution; and $I_f \sim p_{data}(I_f)$ and $I_s \sim p_{data}(I_s)$ indicate that I_f and I_s are selected, respectively from the data distribution p_{data} over the shadow-free and shadow datasets. However, if we use the adversarial loss alone to optimize the generator, there may exist some artifacts on the generated images [15], which may also successfully fool the discriminator [43]. Hence, we take another generator G_s to transform the generated shadow-free image back to its original shadow image and encourage their contents to be the same.

As presented earlier in the introduction, we can produce multiple shadow images from one shadow-free image by adding the shadow regions of different shapes at different image locations. To preserve the consistency between the generated shadow image and the original one, we use a shadow mask M_l as the guidance to indicate the shadow regions, and concatenate the shadow mask M_l with the generated shadow-free image \tilde{I}_f as the input to the generator G_s , which produces the shadow image \tilde{I}_s :

$$\tilde{I}_s = G_s(\tilde{I}_f, M_l) \quad (3)$$

where the shadow mask M_l is the difference between the real shadow image I_s and the generated shadow-free image \tilde{I}_f . The shadow mask is a binary map, where zeros indicate non-shadow regions and ones indicate shadow regions; see Section 3.4 for the details. Then, we formulate the following shadow cycle consistency loss to encourage the reconstructed image \tilde{I}_s to be similar to the original input real shadow image I_s , and optimize the mapping functions in

G_s and G_f by a cycle-consistency constraint:

$$L_{cycle}^a(G_f, G_s) = \mathcal{E}_{I_s \sim p_{data}(I_s)} [\|G_s(G_f(I_s), M_t) - I_s\|_1] . \quad (4)$$

By leveraging the L_1 loss $\|\cdot\|_1$ to calculate the difference on each pixel, generator G_s will learn to produce a shadow image as well as to capture the relationship between the shadow image and shadow mask, *i.e.*, zeros denote the non-shadow regions, while ones denote the shadow regions.

See again the Figure 2(a), we further use a mask M_n with all zero values and the real shadow image I_s as the input of G_s , and generate an image \tilde{I}_s^i , which contains no newly added shadows:

$$\tilde{I}_s^i = G_s(I_s, M_n) . \quad (5)$$

Then, we leverage the shadow identity loss [30] to regularize the output to be close to the input shadow image:

$$L_{identity}^a(G_s) = \mathcal{E}_{I_s \sim p_{data}(I_s)} [\|G_s(I_s, M_n) - I_s\|_1] . \quad (6)$$

Hence, we can encourage that no shadows will be added on the input shadow image in the generated image \tilde{I}_s^i under the guidance of M_n , and we can also preserve the color composition between the input and output images [43].

3.2. Learning from Shadow-free Images

Figure 2(b) shows the framework on how to learn from the shadow-free images for shadow removal. Given a real shadow-free image I_f , we use a generator G_s to produce the shadow image \tilde{I}_s , which is used to fool the discriminator D_s , and makes it hard to distinguish whether it is a real shadow image or not. As mentioned before, for the generator G_s , we need a shadow mask as the input to indicate the shadow regions. Here, we are able to use a mask M_r with any forms of shadows as the guidance and produce the generated shadow image \tilde{I}_s :

$$\tilde{I}_s = G_s(I_f, M_r), \quad D_s(\tilde{I}_s) = \text{real or fake ?} \quad (7)$$

To make the generated shadow regions look real, we randomly select one shadow mask learned from the real shadow image; see Section 3.4 for the details. By leveraging different shadow masks as the guidance, we produce multiple shadow images with different forms of shadows. Therefore, *a large number of shadow images will be created, thus increasing the generalization capability of the deep models.* Finally, we use the adversarial loss to optimize the generator G_s and discriminator D_s :

$$L_{GAN}^b(G_s, D_s) = \mathcal{E}_{I_s \sim p_{data}(I_s)} [\log(D_s(I_s))] + \mathcal{E}_{I_f \sim p_{data}(I_f)} [\log(1 - D_s(G_s(I_f, M_r)))] . \quad (8)$$

To leverage the cycle-consistency constraint, we adopt the generator G_f to produce the shadow-free image \tilde{I}_f from

the generated shadow image \tilde{I}_s , s.t., $\tilde{I}_f = G_f(\tilde{I}_s)$, and use the shadow-free cycle consistency loss to optimize the networks:

$$L_{cycle}^b(G_s, G_f) = \mathcal{E}_{I_f \sim p_{data}(I_f)} [\|G_f(G_s(I_f, M_r)) - I_f\|_1] . \quad (9)$$

Last, we adopt the generator G_f to produce a shadow-free image \tilde{I}_f^i by taking the real shadow-free image I_f as the input, s.t., $\tilde{I}_f^i = G_f(I_f)$, and then use the shadow-free identity loss to force the content of the input and output images to be the same:

$$L_{identity}^b(G_f) = \mathcal{E}_{I_f \sim p_{data}(I_f)} [\|G_f(I_f) - I_f\|_1] . \quad (10)$$

By using the identity loss as a constraint, the generator G_f will learn to remove shadows without changing colors on non-shadow regions.

3.3. Loss Function

In summary, the final loss function for our Mask-ShadowGAN is a weighted sum of the adversarial loss, cycle consistency loss, and identity loss in two parts of the framework:

$$\begin{aligned} L_{final}(G_s, G_f, D_s, D_f) &= \omega_1 (L_{GAN}^a(G_f, D_f) + L_{GAN}^b(G_s, D_s)) \\ &+ \omega_2 (L_{cycle}^a(G_f, G_s) + L_{cycle}^b(G_s, G_f)) \\ &+ \omega_3 (L_{identity}^a(G_s) + L_{identity}^b(G_f)) . \end{aligned} \quad (11)$$

We follow [43] and empirically set ω_1 , ω_2 , and ω_3 as 1, 10, and 5, respectively. Finally, we optimize the whole framework in a minimax manner:

$$\arg \min_{G_s, G_f} \max_{D_s, D_f} L_{final}(G_s, G_f, D_s, D_f) . \quad (12)$$

3.4. Mask Generation

As described earlier, we design a shadow mask to indicate how to generate shadows on the shadow-free images. We obtain the shadow mask M by calculating the difference between the real shadow image I_s and the generated shadow-free image \tilde{I}_f , and then binarizing the result:

$$M = \mathbb{B}(\tilde{I}_f - I_s, t) , \quad (13)$$

where \mathbb{B} indicates the binarization operation, which sets the pixels as one, when their values are greater than the threshold t , otherwise, as zero. We obtain the threshold t by Otsu's algorithm [25], which calculates the optimum threshold to separate shadow and non-shadow regions by minimizing the intra-class variance.

Since we obtain one shadow mask from one real shadow image, we adopt a list to save multiple shadow masks produced from the shadow images in one dataset. During the

training process, the quality of shadow masks increases with the quality of generated shadow-free images \tilde{I}_f . Hence, we update the list of shadow masks by pushing the newly generated mask (high quality) and removing the least recently added mask (low quality). This process is achieved by the *Queue* dataset structure, which obeys the rule of “first in, first out”. Moreover, we empirically set the length of the list as a quarter of image numbers in the real shadow dataset.

When accessing the shadow masks, we set M_l in Figure 2(a) as the newly generated mask from its input shadow image, and randomly choose one mask from the list as M_r in Figure 2(b).

3.5. Network Architecture and Training Strategy

Network architecture. We take the network architecture designed by Johnson *et al.* [16] as our generator network, which includes three convolution operations, followed by nine residual blocks with the stride-two convolutions and two deconvolutions for feature map upsampling. In this network, instance normalization [31] is used after each convolution and deconvolution operation. The generator G_f takes the shadow image with the channel number of three as the input, while the generator G_s adopts the concatenation of the shadow-free image and shadow mask as the input, which has four channels in total. Both G_f and G_s produce a residual image with three channels, which is added with the input image as the final shadow-free or shadow image. For the discriminator D_f and D_s , we use the PatchGAN [15] to distinguish whether the image patches are real or fake.

Training strategy. We initialized the parameters in all generators and discriminators by random noise, which follows a zero-mean Gaussian distribution with standard deviation set as 0.02. Moreover, Adam [20] was used to optimize our networks with the first and second momentum values set as 0.5 and 0.999, respectively. We empirically set the basic learning rate as 2×10^{-4} for the first 100 epochs, gradually reduced it to zero with a linear decay rate in the next 100 epochs, and then stopped the learning. Lastly, we built our model on PyTorch with a mini-batch size of one, and randomly cropped images for data augmentation.

4. Unpaired Shadow Removal Dataset - USR

Existing shadow removal datasets [27, 35] are paired. Typically, we have to fix the camera, take a photo with shadows, then take another one without shadows by removing the associated objects. Due to the varied environment light and camera exposure, a training pair may have inconsistent colors and luminosity; see examples in Figure 3. Also, paired data is available only for limited scenes, thus affecting the generality and practicality of the trained models.

We prepared an unpaired shadow removal dataset named USR with 2,445 shadow images and 1,770 shadow-free

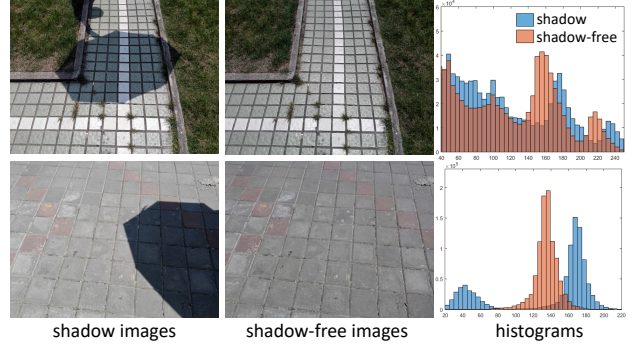


Figure 3: Typical paired shadow and shadow-free images from [35]; note the color inconsistencies revealed by the intensity distributions in the histograms.

images. The dataset contains a large variety of scenes with shadows cast by various kinds of objects, *e.g.*, trees, buildings, traffic signs, persons, umbrellas, railings, *etc.* Very importantly, existing datasets typically cover only hundreds of different scenes (even with thousands of image samples), while ours cover *over a thousand different scenes*. Furthermore, we divide the shadow images in the dataset randomly into 1,956 images for training and 489 for testing, and use all the 1,770 shadow-free images for training (since they are not used in shadow removal testing).

5. Experimental Results

5.1. Datasets and Evaluation Metrics

Datasets. Besides the USR dataset, we employed two recent shadow removal datasets (SRD [27] and ISTD [35]), which contain the paired shadow/shadow-free images and are used for training the existing shadow removal methods.

Evaluation metrics. We followed recent works [13, 27, 35] to evaluate shadow removal performance by computing the root-mean-square error (RMSE) between the ground truth and predicted shadow-free images in LAB color space. In general, a small RMSE indicates a better performance.

5.2. Comparison using USR

First of all, we compare Mask-ShadowGAN with state-of-the-art shadow removal methods on the USR dataset. The purpose here is to show that by leveraging unpaired data, we are able to train a network to learn to remove shadows of more variety shapes for a wider range of scenes.

USR training set. First, we trained our model on the USR training set and applied it to produce shadow-free images on the USR testing set. Also, we applied several state-of-the-art methods to remove shadows on the USR testing set: DSC [13], Gong *et al.* [8], and Guo *et al.* [11]. For DSC, we adopted its public implementation and trained its network on the SRD and ISTD datasets: “DSC-S” and “DSC-

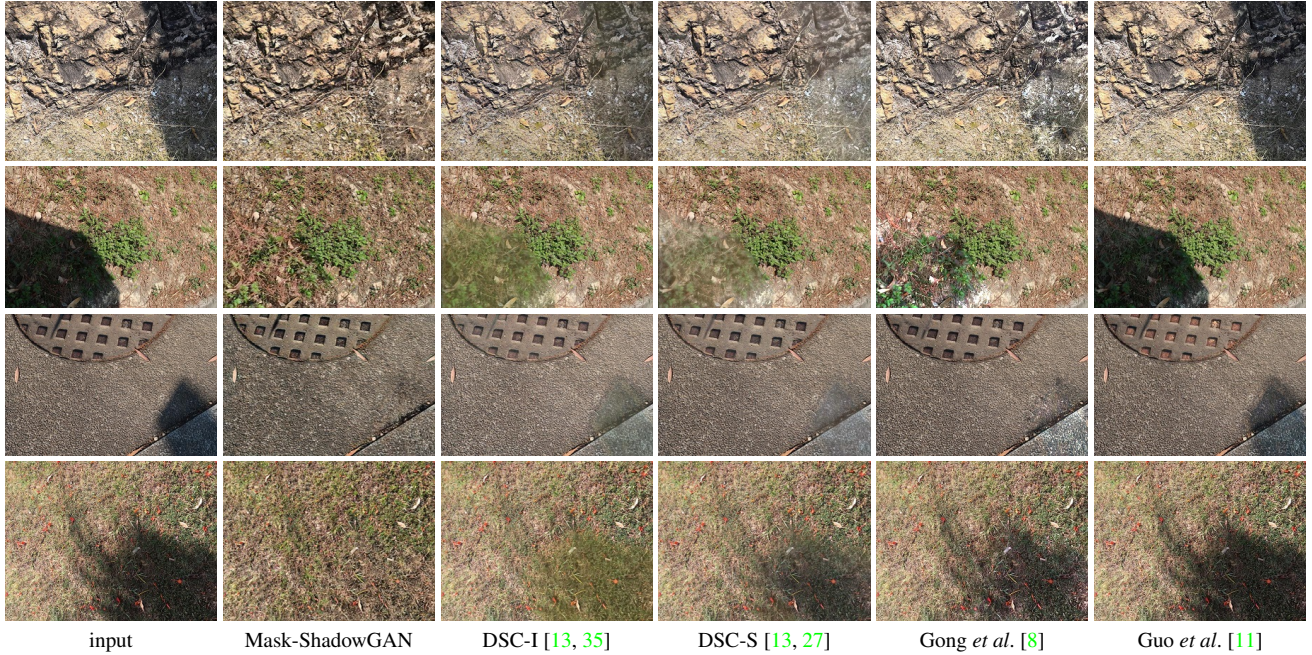


Figure 4: Comparing shadow removal results produced by various methods on the USR dataset.

Table 1: User study results on the USR testing set. Mean ratings (from 1 (bad) to 10 (good)) given by the participants on the shadow removal results.

Methods	Rating (mean & standard dev.)
Mask-ShadowGAN	6.30 ± 2.97
DSC-I [13, 35]	4.78 ± 2.92
DSC-S [13, 27]	4.60 ± 2.66
Gong <i>et al.</i> [8]	2.82 ± 1.76
Guo <i>et al.</i> [11]	2.31 ± 1.90

I” denote the model trained on the SRD dataset and on the ISTD dataset, respectively. Since DSC requires paired shadow and shadow-free images, we cannot re-train it on the USR dataset, which is unpaired. For the other methods, Gong *et al.* and Guo *et al.*, we downloaded and leveraged their public code with recommended parameters to produce the shadow-free image results. Note that the code of STCGAN [35] and DshadowNet [27] are not publicly available, we cannot evaluate them on the USR testing data.

As the unpaired USR dataset has no ground truths, we conducted a user study to evaluate the shadow removal results. First, we generate shadow-free images using Mask-ShadowGAN, as well as using DSC-I, DSC-S, Gong *et al.*, and Guo *et al.*, on the USR testing set (only shadow images). Here, we recruited ten participants: six females and four males, aged 23 to 30 with mean 26.1. For each participant, we randomly selected 150 shadow-free image results (30 per method), presented the results in random order to the participant, and asked the participant to rate the result in a scale from 1 (bad) to 10 (good). Therefore, we obtained 300 ratings (10 participants \times 30 images) per method.

Table 2: User study: participant ratings from 1 (bad) to 10 (good). Different trained models tested on the USR test set.

Trained models	Training set	Rating (mean & standard dev.)
Ours-I	ISTD	4.07 ± 2.93
DSC-I	ISTD	2.38 ± 2.12
Ours-S	SRD	3.38 ± 2.42
DSC-S	SRD	2.93 ± 2.39

Table 1 shows the results. Mask-ShadowGAN received the highest ratings compared to other methods, showing its effectiveness to remove shadows for more diverse scenes, even it was trained just on unpaired data. Further, we performed a statistical analysis on the ratings by conducting t-tests between Mask-ShadowGAN and other methods. All the t-test results show that our results are statistically significant (with $p < 0.001$) than the others, evidencing that the participants prefer our results more than those produced by other methods. Figure 4 shows the visual comparisons, where Mask-ShadowGAN can more effectively remove the shadows and recover the background, while others may blur the images or fail to remove portions of the shadows. Very importantly, our method was trained just on unpaired data.

SRD & ISTD training sets. Also, we trained our method separately on the training sets of SRD (“Ours-S”) and ISTD (“Ours-I”). Then, we applied these models, as well as DSC-S and DSC-I to the USR test set, conducted another user study with four females & six males (aged 22 to 30), and showed 150 randomly-selected shadow-free image results (30 from each trained model) to each participant in random order. Table 2 shows the results. Our method still outperforms the state-of-the-art (DSC), even it was trained on the

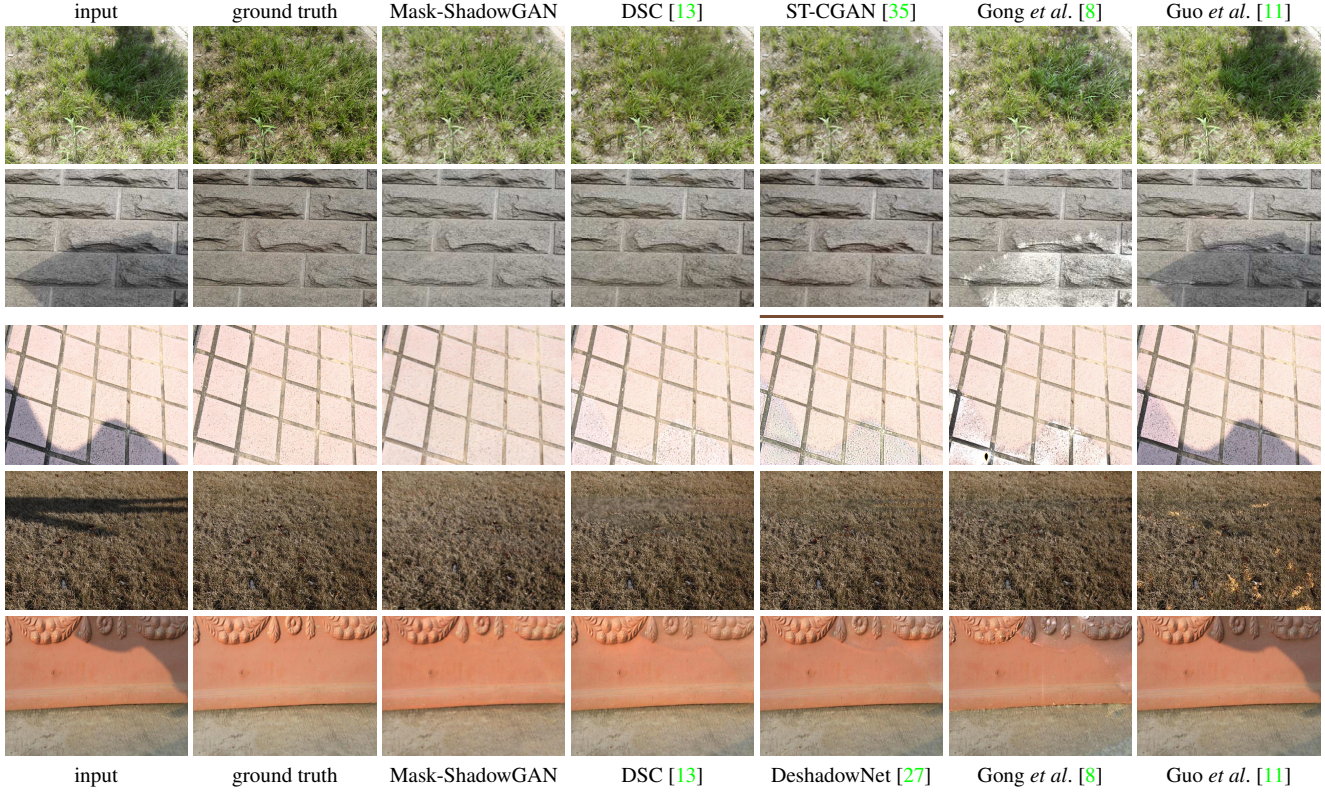


Figure 5: Comparing shadow removal results on the ISTD [35] (top two rows) & SRD [27] (last three rows) datasets. In the fifth column, the top two results are produced from ST-CGAN [35] while the others are produced from DeshadowNet [27].

same training set, since training in an unpaired manner improves the generalization capability of our model.

5.3. Comparison using SRD and ISTD

Next, we compare our method with others on the SRD and ISTD dataset (paired data) using their ground truths. Here, we trained our Mask-ShadowGAN on the SRD training set and tested it on the SRD testing set, then re-trained our model on the ISTD training set and tested the trained model on the ISTD testing set. Since Mask-ShadowGAN is designed to train on unpaired data, we randomly chose an image from the shadow image set and the other from the shadow-free image set per mini-batch during the training.

Quantitative and visual comparison. To compare with other shadow removal methods (DSC [13], ST-CGAN [35], DeshadowNet [27], Gong *et al.* [8], Guo *et al.* [11], Yang *et al.* [40]), we obtained their results directly from the authors or by generating them using the public code with the recommended parameter setting. Here, DSC, ST-CGAN and DeshadowNet are deep networks that produce shadow-free images end-to-end, and their networks were trained on paired shadow and shadow-free images. Gong *et al.*, Guo *et al.*, and Yang *et al.* leverage image priors to remove shadows.

We present the results in Table 3, where our method achieves RMSE values (even trained in an unpaired man-

Table 3: Comparison with the state-of-the-art methods on the SRD [27] and ISTD [35] datasets in terms of RMSE. Note that the code of ST-CGAN [35] and DeshadowNet [27] is not publicly available, so we directly compare with their results on their respective datasets.

Training data	Methods	SRD [27]	ISTD [35]
unpaired	Mask-ShadowGAN	7.32	7.61
	CycleGAN [43]	9.14	8.16
paired	DSC [13]	6.21	6.67
	ST-CGAN [35]	-	7.47
	DeshadowNet [27]	6.64	-
-	Gong <i>et al.</i> [8]	8.73	8.53
	Guo <i>et al.</i> [11]	12.60	9.30
	Yang <i>et al.</i> [40]	22.57	15.63

ner) that are comparable with those of the other deep neural networks trained on paired images, and clearly outperforms the methods based on hand-crafted features. Note also that the code of ST-CGAN [35] and DeshadowNet [27] is not publicly available, and we can only report their results on the datasets used in their published papers.

Figure 5 shows the visual comparison results on these two datasets, which present some challenging cases, e.g., large shadow regions (the first three rows) and shadows across the backgrounds with complex textures (the first, fourth and fifth rows). Although the RMSE value of Mask-

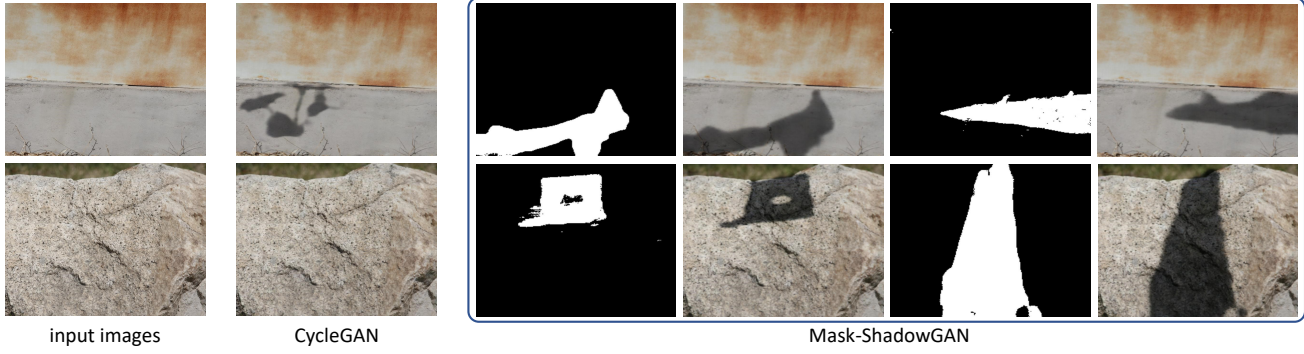


Figure 6: Comparing with CycleGAN [43] on generating shadow images. Note that CycleGAN produces the same output for the same input, while Mask-ShadowGAN can produce different outputs, as guided by the shadow masks (binary images).

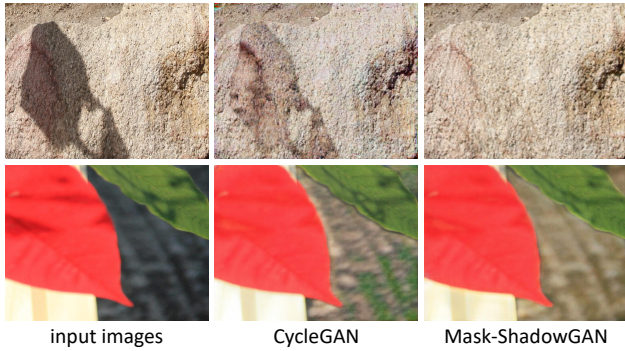


Figure 7: Visual comparison with CycleGAN [43] on generating shadow-free images.

ShadowGAN on these datasets is higher than the deep networks trained on paired data, Mask-ShadowGAN can generate more realistic images and better preserve the texture details occluded by shadows; see again the first, fourth and fifth rows in Figure 5. This is because we learn to remove shadows from the reliable intrinsic statistics of real shadow/shadow-free images, and avoid the unrealistic grayish (blurry) outputs through adversarial learning.

Comparison with CycleGAN. Further, we compared our method with CycleGAN [43], which is designed for general image-to-image translations using unpaired training data. Here, we adopted the author-provided implementations and used the same parameter setting as our Mask-ShadowGAN to re-train the model on the SRD and ISTD training sets.

Table 3 reports the results, showing that our method outperforms CycleGAN on both datasets. By leveraging the shadow masks to guide the shadow generation for both real and generated shadow-free images, we can effectively provide constraints to the network to guide its exploration in the search space for producing more realistic results. Figure 7 shows visual comparison results. Mask-ShadowGAN can clearly remove the shadows, but CycleGAN tends to produce artifacts on the regions occluded by the shadows. Besides, Figure 6 shows shadow images generated from real shadow-free images by Mask-ShadowGAN and by Cycle-

GAN. While CycleGAN always produces the same shadow image for the same shadow-free input, Mask-ShadowGAN is able to produce multiple realistic shadow images with the help of the shadow masks, which are also learned in the network automatically from some real shadow images.

6. Conclusion

In this work, we present a novel generative adversarial framework, named as Mask-ShadowGAN, for shadow removal based on unpaired shadow and shadow-free images. Our key idea is to transform the uncertain shadow-free-to-shadow image translation into a deterministic image translation with the guidance of the shadow masks, which are learned from the real shadow images automatically. Further, we construct the first unpaired shadow removal (USR) dataset, test our method on various datasets, and compare it with the state-of-the-art methods to show its quality, both quantitatively and visually. However, our method assumes a small domain difference (*e.g.*, variations in scene contents) between the unpaired shadow and shadow-free images. Also, to aim for better results, we generally need a larger training set. In the future, we plan to use the generated shadow masks to facilitate new applications, such that shadow editing, where we may manipulate shadows rather than just removing them. We also plan to explore the mask generation technique for other image translation applications, where mapping between image sets is not one-to-one, *e.g.*, learn to remove/synthesize rain or snow from unpaired real rain/snow and rain-free/snow-free images.

Acknowledgments. This work was supported by the National Basic Program of China, 973 Program (Project no. 2015CB351706), the Shenzhen Science and Technology Program (JCYJ20170413162256793 & JCYJ20170413162617606), the Research Grants Council of the Hong Kong Special Administrative Region (Project no. CUHK 14201918), and the CUHK Research Committee Funding (Direct Grants) under project code - 4055103. Xiaowei Hu is funded by the Hong Kong Ph.D. Fellowship.

References

- [1] Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, and Aaron C. Courville. Augmented CycleGAN: Learning many-to-many mappings from unpaired data. In *ICML*, pages 195–204, 2018. 3
- [2] Eli Arbel and Hagit Hel-Or. Shadow removal using intensity surfaces and texture anchor points. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1202–1216, 2011. 2
- [3] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015. 2
- [4] Graham D. Finlayson, Mark S. Drew, and Cheng Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85(1):35–57, 2009. 2
- [5] Graham D. Finlayson, Steven D. Hordley, and Mark S. Drew. Removing shadows from images. In *ECCV*, pages 823–836, 2002. 2
- [6] Graham D. Finlayson, Steven D. Hordley, Cheng Lu, and Mark S. Drew. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):59–68, 2006. 2
- [7] Clément Fredembach and Graham Finlayson. Hamiltonian path-based shadow removal. In *BMVC*, volume 2, pages 502–511, 2005. 2
- [8] Han Gong and D. P. Cosker. Interactive shadow removal and ground truth for variable scene categories. In *BMVC*, pages 1–11, 2014. 2, 5, 6, 7
- [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*, pages 2672–2680, 2014. 2
- [10] Maciej Gryka, Michael Terry, and Gabriel J. Brostow. Learning to remove soft shadows. *ACM Transactions on Graphics*, 34(5):153, 2015. 1, 2
- [11] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2956–2967, 2013. 1, 2, 5, 6, 7
- [12] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. 2
- [13] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. to appear. 1, 2, 5, 6, 7
- [14] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *CVPR*, pages 7454–7462, 2018. 2
- [15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 2, 3, 5
- [16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016. 5
- [17] Salman Hameed Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Automatic feature learning for robust shadow detection. In *CVPR*, pages 1939–1946, 2014. 2
- [18] Salman Hameed Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Automatic shadow detection and removal from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):431–446, 2016. 1, 2
- [19] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, pages 1857–1865, 2017. 2
- [20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [21] Hieu Le, Tomas F. Yago Vicente, Vu Nguyen, Minh Hoai, and Dimitris Samaras. A+D Net: Training a shadow detector with adversarial shadow attenuation. In *ECCV*, pages 662–678, 2018. 2
- [22] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, pages 35–51, 2018. 3
- [23] Feng Liu and Michael Gleicher. Texture-consistent shadow removal. In *ECCV*, pages 437–450, 2008. 2
- [24] Ankit Mohan, Jack Tumblin, and Prasun Choudhury. Editing soft shadows in a digital photograph. *IEEE Computer Graphics and Applications*, 27(2):23–31, 2007. 2
- [25] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. 4
- [26] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *CVPR*, pages 2701–2710, 2017. 2
- [27] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W.H. Lau. DeshadowNet: A multi-context embedding deep network for shadow removal. In *CVPR*, pages 4067–4075, 2017. 1, 2, 5, 6, 7
- [28] Yael Shor and Dani Lischinski. The shadow meets the mask: Pyramid-based shadow removal. In *Computer Graphics Forum*, volume 27, pages 577–586. Wiley Online Library, 2008. 2
- [29] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, pages 2107–2116, 2017. 2
- [30] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 4
- [31] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 5
- [32] Tomás F. Yago Vicente, Minh Hoai, and Dimitris Samaras. Leave-one-out kernel optimization for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):682–695, 2018. 1, 2

- [33] Tomás F. Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *ECCV*, pages 816–832, 2016. 2
- [34] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103, 2008. 2
- [35] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, pages 1788–1797, 2018. 1, 2, 5, 6, 7
- [36] Xiaolong Wang, Kaiming He, and Abhinav Gupta. Transitive invariance for self-supervised visual representation learning. In *ICCV*, pages 1329–1338, 2017. 2
- [37] Tai-Pang Wu and Chi-Keung Tang. A bayesian approach for shadow extraction from a single image. In *ICCV*, volume 1, pages 480–487, 2005. 2
- [38] Tai-Pang Wu, Chi-Keung Tang, Michael S. Brown, and Heung-Yeung Shum. Natural shadow matting. *ACM Transactions on Graphics (TOG)*, 26(2):8, 2007. 2
- [39] Chunxia Xiao, Ruiyun She, Donglin Xiao, and Kwan-Liu Ma. Fast shadow removal using adaptive multi-scale illumination transfer. In *Computer Graphics Forum*, volume 32, pages 207–218. Wiley Online Library, 2013. 2
- [40] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. *IEEE Transactions on Image Processing*, 21(10):4361–4368, 2012. 2, 7
- [41] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dual-GAN: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2849–2857, 2017. 2
- [42] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson W.H. Lau. Distraction-aware shadow detection. In *CVPR*, pages 5167–5176, 2019. 2
- [43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 2, 3, 4, 7, 8
- [44] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*, pages 121–136, 2018. 2