From Shadow Generation to Shadow Removal

Zhihao Liu[1], Hui Yin[1,*], Xinyi Wu[2], Zhenyao Wu[2], Yang Mi[3], Song Wang[2,*]
[1]Beijing Jiaotong University, China [2]University of South Carolina, USA [3]China Agriculture University, China
{16120394,hyin}@bjtu.edu.cn,{xinyiw,zhenyao}@email.sc.edu,miy@cau.edu.cn,songwang@cec.sc.edu

Abstract

Shadow removal is a computer-vision task that aims to restore the image content in shadow regions. While almost all recent shadow-removal methods require shadow-free images for training, in ECCV 2020 Le and Samaras introduces an innovative approach without this requirement by cropping patches with and without shadows from shadow images as training samples. However, it is still laborious and time-consuming to construct a large amount of such unpaired patches. In this paper, we propose a new G2R-ShadowNet which leverages shadow generation for weakly-supervised shadow removal by only using a set of shadow images and their corresponding shadow masks for training. The proposed G2R-ShadowNet consists of three sub-networks for shadow generation, shadow removal and refinement, respectively and they are jointly trained in an end-to-end fashion. In particular, the shadow generation sub-net stylises non-shadow regions to be shadow ones, leading to paired data for training the shadow-removal sub-net. Extensive experiments on the ISTD dataset and the Video Shadow Removal dataset show that the proposed G2R-ShadowNet achieves competitive performances against the current state of the arts and outperforms Le and Samaras' patch-based shadow-removal method.

1. Introduction

Shadows are areas of darkness in a scene where the light is fully or partially occluded. Shadows are very common in natural images and might bring challenges to many existing computer vision tasks [24, 12, 18, 14]. Shadow removal by restoring the image information in shadow regions have been a long studied research problem [2, 15, 27, 19, 40, 23] and has been shown to be beneficial to improve the performance in various tasks.

Recently, with the use of the convolutional neural networks (CNNs), many learning based shadow removal approaches [25, 32, 9, 10, 16, 22, 20, 17] have been proposed, resulting in significantly better performance than the traditional ones [4, 8, 13, 38]. For most of them [25, 32, 9, 16] a set of paired shadow images and their corresponding shadow-free version are used to train the network in a fully-supervised manner. However, it is difficult to capture and collect such paired image data in uncontrolled natural environment due to the illumination change from time to time. If we capture such data pairs in a controlled lab environment, limited scenarios might also weaken the generalisation ability of the trained model.

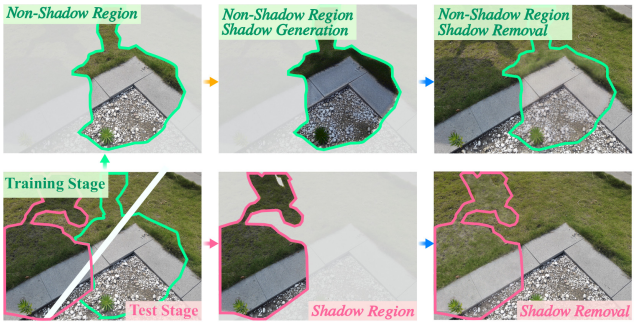

Figure 1. An illustration of our basic idea of incorporating shadow generation for learning shadow removal by using only shadow images. The pixels located in the pink and green boundaries of the input shadow image form the shadow and random non-shadow regions, respectively. The pink and green arrows stand for the masking operation that only preserves the region with the value of 1 on the mask, while the orange and blue arrows represent the process of shadow generation and shadow removal, respectively.

To address these problems, recent researches [10, 22] start to explore unsupervised methods for shadow removal by using unpaired shadow and shadow-free images. However, these unsupervised methods might introduce huge domain gap between the shadow and shadow-free images in the training set. In addition, in practice, it is still difficult to capture a large set of shadow-free images with good variety. To handle this problem, in their ECCV20 paper [17], Le and Samaras introduces a novel unsupervised shadow removal method using only shadow images. More specifically, they leverage the fact that a shadow image usually contains both shadow and non-shadow regions. This way, a set of shadow and shadow-free patches can be cropped to

*Co-corresponding authors. Code is available at https://github.com/hhqweasd/G2R-ShadowNet.

construct unpaired data for network training. By cropping unpaired patches from the same images, their domain gap can also be well controlled.

Although the patch cropping in [17] only requires the shadow masks that can be obtained by using an existing shadow detection method [43, 18, 41, 9, 33], it involves a careful design of cropping-window size, a set of strict physics-based constraints, and heavy computational load [17]. In this paper, we propose a new approach to address these problems by incorporating a shadow generation module, while keeping the desirable property of using only shadow images, as illustrated in Fig.1.

Specifically, we propose a new G2R-ShadowNet that consists of three sub-networks for shadow generation, shadow removal and refinement, respectively. Given an input shadow image, the shadow-generation sub-net generates pseudo shadows for each shadow-free region and such pseudo shadows are then paired with the corresponding original shadow-free region to form the training data. After that, these constructed pair data are used to train the shadow removal sub-net to remove the generated shadows. Finally, the shadow-removal results are refined by leveraging the context information such that their colour and illumination are consistent with their surrounding areas. We conduct extensive experiments on the ISTD dataset and the Video Shadow Removal dataset to demonstrate the effectiveness of our proposed method.

The main contributions of this work are as follows:

- We tackle the shadow removal task from a novel perspective of constructing paired shadow and non-shadow data using only the shadow images and the corresponding shadow masks.
- We develop G2R-ShadowNet, a novel shadow-removal network, which consists of three sub-nets for shadow generation, shadow removal, and refinement, respectively. G2R-ShadowNet is weakly-supervisedly trained in an end-to-end fashion.
- We conduct extensive experiments on two public datasets and show that the proposed G2R-ShadowNet achieves competitive performances against the current state of the arts and outperforms Le and Samaras' patch-based shadow-removal method.

2. Related Work

2.1. Shadow generation

Our proposed G2R-ShadowNet contains a shadow generator which employs the generative adversarial networks (GAN) [6] for generating shadows on shadow-free regions. GAN-based shadow generation has been studied by many researchers. Zhang *et al*. [39] proposed a GAN to synthesise shadows for virtual objects that are inserted into images and train the network with shadow masks and paired shadow/shadow-free images. Similarly, Liu *et al*. [21] developed a ARShadowGAN for augmented reality in single light scenes, which exploits attention mechanism to model the mapping relationship between the shadow of the virtual objects and the real-world environment. These two works [39, 21] rely on fully-supervised training which requires the shadow and shadow-free images as well as the shadow masks. In [10, 22], a mapping is learned between the unpaired shadow and shadow-free images for shadow removal, in which shadow generators are trained to match the distributions of the generated shadows and the real shadows based on unpaired images. However, the shadow-free images are the prerequisites for all the above methods to generate the shadow images. In contrast, in this paper we sample both shadow and shadow-free regions only from shadow images, which is also different from other methods that need images from the target domain for their respective applications [34, 1] or generate simulated data on the whole images from the source domain [36].

2.2. Shadow removal

Traditional approaches remove shadows according to image gradients [4, 7], illumination information [26, 37, 35, 38], and region properties [8, 30]. Recently, supervised learning based methods trained with large-scale paired datasets boost the shadow-removal performance significantly [25, 32, 9, 3, 16, 20]. However, as mentioned above, paired shadow and shadow-free images are difficult to obtain in practice. To get rid of the dependence on paired data, Hu *et al*. [10] proposed a Mask-ShadowGAN framework based on the CycleGAN [42], which leverages unpaired data to learn the adaptation from the shadow-free domain to the shadow domain and vice versa. Liu *et al*. [22] later developed a LG-ShadowNet framework to improve the Mask-ShadowGAN [10] by introducing a lightness-guided strategy, which uses the learned lightness features to guide the learning of shadow removal.

However, all these methods still need shadow-free images for training which requires very strict acquisition conditions and may introduce huge gap between the source (non-shadow) domain and the target (shadow) domain. In this paper, the shadow and non-shadow regions are sampled from the same shadow image and therefore, the distribution difference between the two domains is much smaller. In addition, we apply the adversarial training only in the shadow-generation sub-net, while Mask-ShadowGAN and LG-ShadowNet apply it in both shadow-generation and shadow-removal networks, which makes the whole framework more difficult to converge.

As mentioned above, most related to our work is [17], where unpaired data in form of patches are cropped from the same shadow image according to the shadow mask. Therefore, the constructed data show small domain gap. How-

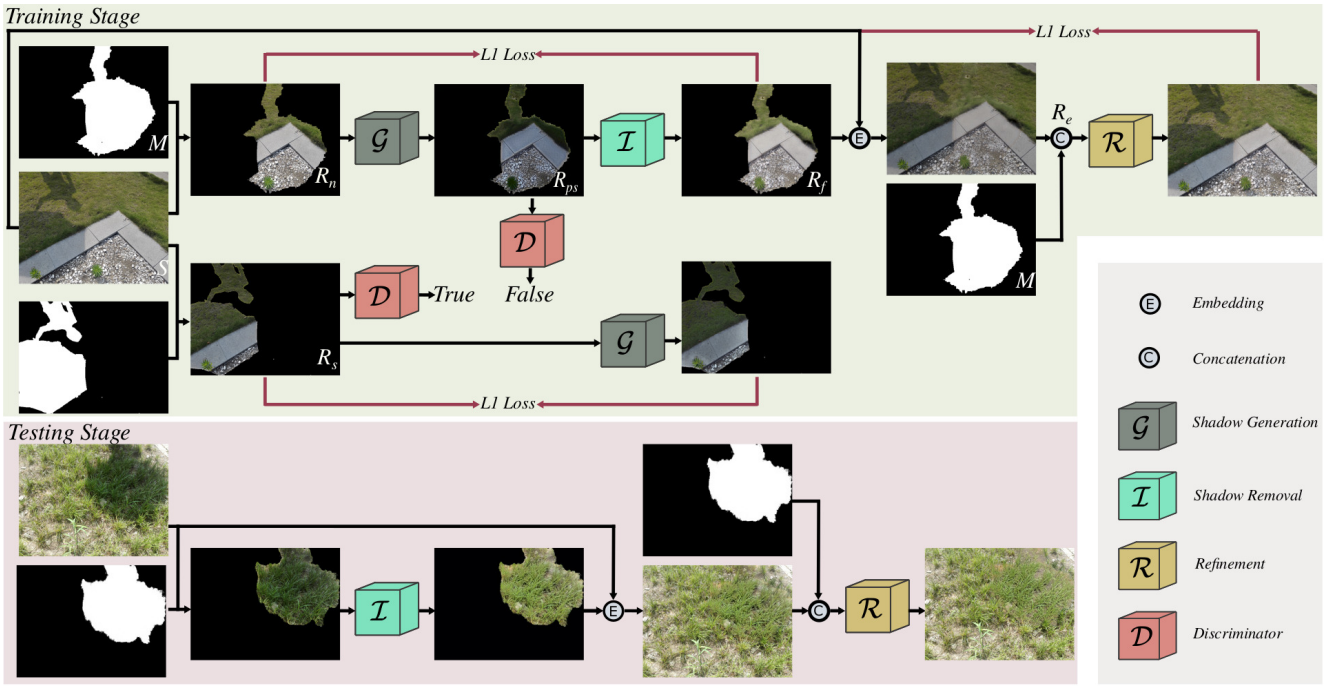

Figure 2. The network architecture of our proposed G2R-ShadowNet. It takes in a shadow image and its corresponding shadow mask to produce the shadow-free result in the shadow regions. The training stage involves all the three sub-nets of shadow generation, shadow removal and refinement, while the testing stage does not perform shadow generation.

ever, this patch-based method suffers from a heavy computational load because of the repetitive cropping of a small step size. In addition, the strict physics-based constraints used in this method limits the shadow types that can be handled. Our proposed G2R-ShadowNet can address these issues by constructing paired data from the same shadow image using the shadow mask without repetitive cropping.

3. Methodology

In this section, we elaborate on the overall network architecture of our proposed G2R-ShadowNet, which includes three modules: shadow generation, shadow removal, and refinement. All the three parts are jointly trained in an end-to-end fashion, as illustrated in Fig. 2.

3.1. Shadow generation sub-net

The shadow generation sub-net is used to construct our training data, i.e., paired shadow and non-shadow regions, for the shadow removal network. Specifically, we first crop the shadow region R_s from the input image S by applying the corresponding shadow mask: the remaining area of the image is setting to 0 automatically. Then we randomly pick another shadow mask M from the masks of the training set and apply it on the shadow-free region of S to obtain a non-shadow region R_n, whose area approximates the area of the shadow region R_s by the constraint

$$\texttt{Area}(R_n)/\texttt{Area}(R_s) \in (1-\alpha, 1+\alpha), \qquad (1)$$

where $\texttt{Area}(\cdot)$ computes the area of the given region and α is a tolerance value which is set to 0.2 in our experiments. Note that this constraint will be discarded in the cases where R_s covers more than half of the whole image and we randomly select a non-shadow region for masking and shadow generation.

Using the above operations, we construct many pairs of unaligned data from both shadow and non-shadow domains, which are used to train our shadow generator $\mathcal{G}$ to generate pseudo shadow R_{ps} on the non-shadow region of S via adversarial training. A discriminator $\mathcal{D}$ is employed to distinguish the pseudo shadow R_{ps} from a randomly sampled real shadow to help train $\mathcal{G}$ and ensure the data distribution similarity between the two domains.

The architecture of $\mathcal{G}$ mainly follows the generator proposed by Hu *et al.* [10]. It consists of three convolutional layers with a stride of 2 to decrease the resolution of the input image, followed by nine residual blocks to extract features, and ends with three deconvolutional layers to generate the output with the same resolution as S. Besides, the instance normalisation [29] is applied after each convolutional operation. For the architecture of $\mathcal{D}$, we directly employ the one proposed in PatchGAN [11]. All the inputs and outputs of $\mathcal{G}$ and the inputs of $\mathcal{D}$ are 3-channel images in LAB colour space.

The objective functions to train the shadow generator $\mathcal{G}$

and the discriminator $\mathcal{D}$ are defined as:

$$L_{Gen}(\mathcal{G}) = \frac{1}{2}\mathbb{E}_{R_n \sim p(R_n)}\big[(\mathcal{D}(\mathcal{G}(R_n)) - 1)^2\big], \quad (2)$$

$$\begin{aligned} L_{Dis}(\mathcal{D}) =& \frac{1}{2}\mathbb{E}_{R_n \sim p(R_n)}\big[(\mathcal{D}(\mathcal{G}(R_n)))^2\big] \\ &+ \frac{1}{2}\mathbb{E}_{R_s \sim p(R_s)}\big[(\mathcal{D}(R_s) - 1)^2\big]. \end{aligned} \quad (3)$$

The combined loss function for the adversarial training is:

$$L_{GAN} = L_{Gen}(\mathcal{G}) + L_{Dis}(\mathcal{D}). \quad (4)$$

In addition, to ensure that the shadow generation sub-net produces high quality synthetic shadows, real shadow R_s is also fed to $\mathcal{G}$, and we apply the identical loss [28] to encourage $\mathcal{G}$ to generate the same shadow as the input R_s, which is defined as:

$$L_{iden}(\mathcal{G}) = \mathbb{E}_{R_s \sim p(\mathcal{R}_s)}\big[\,\|\mathcal{G}(R_s), R_s\|_1\,\big], \quad (5)$$

where $\|, \|_1$ represents L_1 loss.

3.2. Shadow removal sub-net

We use the pairs of R_{ps} and R_n as the inputs of our shadow removal sub-net for learning to remove the pseudo shadows that are generated by the generator $\mathcal{G}$. Specifically, the shadow removal sub-net $\mathcal{I}$ has the same structure as the generator $\mathcal{G}$, and it takes the output of $\mathcal{G}$, i.e., R_{ps}, as the input to produce a shadow-free result R_f that shares the same content as R_{ps}. The loss function to train $\mathcal{I}$ is defined as:

$$\begin{aligned} L_{rem}(\mathcal{G}, \mathcal{I}) =& \mathbb{E}_{R_{ps} \sim p(\mathcal{R}_{ps})}\big[\,\|\mathcal{I}(R_{ps}), R_n\|_1\,\big] \\ =& \mathbb{E}_{R_n \sim p(\mathcal{R}_n)}\big[\,\|\mathcal{I}(\mathcal{G}(R_n)), R_n\|_1\,\big]. \end{aligned} \quad (6)$$

Note that the gradient computed by this loss will be propagated back to the shadow generator $\mathcal{G}$ through R_{ps} and therefore, it can be regarded as a cycle loss [42] for training both $\mathcal{G}$ and $\mathcal{I}$.

3.3. Refinement sub-net

The output R_f from the shadow removal sub-net is then embedded into the input image S to obtain R_e which is formulated as:

$$R_e = \langle R_f + S - R_n, M \rangle, \quad (7)$$

where $\langle \cdot, \cdot \rangle$ denotes the concatenation operation, and M is the shadow mask used in previous parts which covers the region to be processed. The obtained R_e is a 4-channel tensor including 3 channels for the image and 1 channel for the mask. However, the colour of R_f might not be fully consistent with that of the other regions of image S. To address this problem, we further develop a refinement sub-net $\mathcal{R}$ to refine R_e by exploiting context information of the shadow region over the original whole shadow image.

Specifically, our refinement network $\mathcal{R}$ takes in R_e as input and outputs a refined image R_r. The network $\mathcal{R}$ shares the same structure as both $\mathcal{I}$ and $\mathcal{G}$ except for the number of the input channels. A per-pixel loss between the refined image R_r and the input S is computed to train the refinement network, which is defined as:

$$\begin{aligned} &L_{full}(\mathcal{G}, \mathcal{I}, \mathcal{R}) \\ &\quad = \mathbb{E}_{R_f \sim p(\mathcal{R}_f)}\big[\,\|\mathcal{R}(R_e), S\|_1\,\big]. \end{aligned} \quad (8)$$

This loss function is calculated according to the context information of the shadow region across the whole shadow image and the computed gradient will be propagated back to $\mathcal{G}$ and $\mathcal{I}$.

To further emphasise the content of the output in the same region as R_n to be the same as that part in S, we apply the following loss function:

$$\begin{aligned} &L_{area}(\mathcal{G}, \mathcal{I}, \mathcal{R}) \\ &\quad = \mathbb{E}_{R_f \sim p(\mathcal{R}_f)}\big[\sum_n \psi(M)|\mathcal{R}(R_e) - S|\big], \end{aligned} \quad (9)$$

where n is the number of pixels in the input image S. ψ denotes the image dilation function with a kernel size of τ, which produces a dilated mask to guide the model to pay more attention to the adjacent area of R_n.

3.4. Loss function

By combining all the loss functions proposed for the above three sub-nets, the total loss $\mathcal{L}$ for training the shadow generator $\mathcal{G}$, the shadow removal sub-net $\mathcal{I}$ and the refinement sub-net $\mathcal{R}$ is defined as

$$\begin{aligned} \mathcal{L} =& \omega_1 L_{GAN} + \omega_2 L_{iden} \\ &+ \omega_3 L_{rem} + \omega_4 L_{full} + \omega_5 L_{area}, \end{aligned} \quad (10)$$

where ω_1, ω_2, ω_3, ω_4, and ω_5 are the weights to balance different loss terms and are set to 1.0, 5.0, 1.0, 1.0, and 1.0, respectively in our experiments.

4. Experiments

4.1. Datasets and evaluation metrics

ISTD [32, 16] The ISTD dataset is proposed for both shadow detection and shadow removal and the data are collected under various illumination conditions with different shadow shapes. In total, it contains 1,870 triplets of shadow, shadow mask and shadow-free images with a resolution of 480×640, where 1,330 triplets for training the rest 540 for testing. Following [16], we apply the adjusted testing set with reduced illumination difference between the shadow and shadow-free images in the original dataset. In training, our model uses the shadow images and the corresponding ground-truth shadow masks. In testing, we employ the

shadow detector proposed by Zhu *et al*. [43] to obtain the shadow mask of the shadow images in the test set. The shadow detector is trained on both the training set of ISTD and the SBU [31] datasets, and it achieves 2.4 Balance Error Rate on the testing set of ISTD when evaluated using the ground-truth shadow masks provided by ISTD.

Video Shadow Removal Dataset [17] The Video Shadow Removal dataset contains 8 videos whose contents are static scenes without moving objects. This dataset also provides a corresponding V_{max} image for each video and a moving-shadow mask for each frame. The V_{max} image is obtained by taking the maximum intensity value at each pixel location across the whole video, which is regarded as the shadow-free ground truth of the video. The moving-shadow mask covers the pixels appearing in both the shadow and shadow-free regions of the video, which represents the region for evaluation. We follow the setting in the official code of [17] using a threshold of 80 to obtain the moving-shadow mask. For this data, we apply the shadow detector [43] that is trained only on the SBU dataset to generate the shadow masks for our experiments.

Evaluation metrics For all experiments, we use the Root-Mean-Square Error (RMSE), Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) as the evaluation metrics. Following [32, 10, 16, 22, 17], we compute the RMSE between the produced shadow-free image and the ground-truth images in the LAB colour space. While some recent work [17] computes RMSE at each pixel and then averages the score over all the pixels, we compute RMSE *on each image* and then average the score over all images/frames for both the ISTD dataset and the video dataset. Our computed RMSE emphasise more the quality of each image on shadow and non-shadow regions and is more consistent with other metrics such as PSNR and SSIM. We also compute PSNR and SSIM scores in the RGB colour space to evaluate our method. For RMSE, the lower the better while for PSNR and SSIM, the higher the better.

4.2. Experimental settings

We implement our proposed G2R-ShadowNet using PyTorch with a single NVIDIA GeForce GTX 2080ti GPU. We initialise our model using a Gaussian distribution with a mean of 0 and a standard deviation of 0.02. We employ the Adam optimiser to train our network with the first and the second momentum setting to 0.5 and 0.999, respectively. We train the whole model for 100 epochs and the base learning rate is set to 2×10^{-4} for the first 50 epochs and then we apply a linear decay strategy to decrease it to 0 for the rest epochs. The batch size is set to 1 for all experiments. The size of the dilated kernel τ in Eq. (9) is experimentally set to 50 based on our ablation study on various values. For the data augmentation, we apply the random cropping and random flipping to avoid the over-fitting problem. The random cropping is implemented by first scaling each image to 448×448 and then randomly cropping a 400×400 region from the scaled image.

The network training involves all three sub-nets, and they impact each other through a forward or backward signal flow, e.g., the gradient from $\mathcal{R}$ can be propagated back to $\mathcal{I}$ and $\mathcal{G}$. While in the testing stage, given the shadow image and its shadow mask, only the shadow removal and refinement network are employed to produce the final shadow removal result with a resolution of 256×256 for evaluation. It approximately takes 16 hours to train the proposed G2R-ShadowNet on the ISTD dataset and 0.06 seconds to perform shadow removal for a test image.

4.3. Ablation study

To demonstrate the effectiveness of each key component of the proposed G2R-ShadowNet, we train and test several model variants on ISTD.

We first conduct an experiment to study each design of the refinement sub-net $\mathcal{R}$ by comparing it with the other two variants. One is obtained by removing $\mathcal{R}$ and the loss functions that are related to $\mathcal{R}$. The other one is obtained without using the shadow mask M as the input which means the region that is going to be refined is not known. The quantitative results are reported in Table 1. The results indicate that the sub-net $\mathcal{R}$ plays quite an important role in our whole framework, which significantly improves the quality of the shadow removal result in terms of all three metrics. Including the shadow masks as the input is effective and brings improvements to all the metrics as well.

Table 1. Ablation study to verify the effectiveness of the refinement sub-net of our proposed G2R-ShadowNet on the test set of ISTD using all three evaluation metrics. Hereafter, 'Shadow Region' represents that the metric is computed only on the shadow region of the image, and 'All' represents that the metric is computed on the whole image. The best and the second best results are highlighted with **bold** font and underline, respectively.

Method	Shadow Region			All		
	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM
Ours w/o $\mathcal{R}$	<u>12.3</u>	29.20	0.975	4.6	26.04	0.913
Ours w/o M	12.5	<u>29.68</u>	<u>0.977</u>	<u>4.3</u>	<u>27.49</u>	<u>0.940</u>
Ours	**8.9**	**33.58**	**0.979**	**3.9**	**30.52**	**0.944**

Next, we perform an ablation study to verify the effectiveness of the joint training strategy of the three sub-nets in our method. Basically, the shadow generation, shadow removal, and the refinement of our proposed G2R-ShadowNet can impact the learning of each other through the backward signal flow. Therefore, we try to detach the result of each sub-net individually and train each variants one-by-one to see how one impacts the others. For instance, when the refinement result is detached, the back-propagated signal from the refinement sub-net $\mathcal{R}$ is not passed to $\mathcal{G}$ and $\mathcal{I}$.

Input | w/o L_{GAN} | w/o L_{iden} | w/o L_{rem} | w/o L_{full} | w/o L_{area} | Ours | Ground truth

Figure 3. Visual comparisons for ablation study on the use of each loss term.

The quantitative results are reported in Table 2, from which the effectiveness of our designs are well justified. Particularly, we observe that the joint training of the three sub-nets (the last row) can boost the performance and achieve the highest score on both PSNR and SSIM compared with other variants. Especially, when results of all three sub-nets are detached, the performance drops in the shadow region on all the metrics. By connecting only two of them for training, we can achieve certain performance gains. By comparing the second and third model variants, we find that the connection of $\mathcal{G}$ and $\mathcal{I}$ contributes more to the performance than the connection of $\mathcal{I}$ and $\mathcal{R}$.

Table 2. Ablation study to verify the effectiveness of the joint training strategy of the three sub-nets of our proposed G2R-ShadowNet on ISTD. '$\leftarrow$' and '$\nleftarrow$' in the first column denote the connection and detachment during back-propagation, respectively.

Method	Shadow Region			All		
	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM
Input image	37.0	20.84	0.927	8.5	20.45	0.893
$\mathcal{G} \nleftarrow \mathcal{I} \nleftarrow \mathcal{R}$	9.3	33.43	0.977	3.9	30.24	0.941
$\mathcal{G} \nleftarrow \mathcal{I} \leftarrow \mathcal{R}$	9.0	33.27	0.977	**3.8**	30.36	0.943
$\mathcal{G} \leftarrow \mathcal{I} \nleftarrow \mathcal{R}$	**8.8**	33.46	**0.979**	**3.8**	30.48	**0.944**
$\mathcal{G} \leftarrow \mathcal{I} \leftarrow \mathcal{R}$	8.9	**33.58**	**0.979**	3.9	**30.52**	**0.944**

We also conduct another ablation study to justify the effectiveness of each loss function by training our model without a specific loss term for each time. We report the quantitative results in Table 3. From rows 1-2, we observe that the RMSE performance drops a lot without using L_{GAN} and L_{iden}. The shadow removal loss L_{rem} is also important and brings performance improvement in terms of all the metrics. When L_{rem} is removed from the total loss, the shadow removal sub-net and the refinement sub-net are trained as a whole, which lacks individual constraints. Besides, removing L_{full} leads to performance drop, which verifies the benefit of using the whole image as a constraint to train our G2R-ShadowNet. Finally, when L_{area} is not calculated, the performance also slightly drops in the shadow region. As shown in Fig. 3, the qualitative results are largely consistent with the above quantitative results in justifying the effectiveness of each loss term. Compared with the model training by combining all the loss terms, the other variants that are trained based on a subset of loss terms may cause obvious artefacts on the results, e.g., a white area in the shadow edge as shown in Fig. 3 (column 6).

We also carry out a set of experiments to explore the impact of choosing different dilated kernel size τ in L_{area}. Specifically, we set τ to 0, 5, 15, 50, and 100 and train our model, respectively. Quantitative results are reported in Table 4, which shows that 50 is the optimal dilated kernel size that help achieve the best performance.

Table 3. Ablation study on the choices of the loss functions for the proposed G2R-ShadowNet.

Method	Shadow Region			All		
	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM
Ours w/o L_{GAN}	42.1	19.62	0.911	9.3	19.27	0.878
Ours w/o L_{iden}	43.1	21.57	0.878	9.8	20.70	0.825
Ours w/o L_{rem}	9.8	32.71	0.977	4.1	29.91	0.942
Ours w/o L_{full}	12.3	29.78	0.966	4.3	27.66	0.923
Ours w/o L_{area}	9.3	33.22	**0.979**	**3.8**	30.40	0.943
Ours	**8.9**	**33.58**	**0.979**	3.9	**30.52**	**0.944**

Table 4. Influence of the dilation kernel τ in L_{area} to the performance of the proposed G2R-ShadowNet.

Method	Shadow Region			All		
	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM
$\tau = 0$	9.5	33.01	0.977	**3.9**	30.11	0.939
$\tau = 5$	9.5	32.70	0.976	4.0	29.89	0.939
$\tau = 15$	9.1	33.54	**0.980**	**3.9**	30.40	**0.944**
$\tau = 50$	**8.9**	**33.58**	0.979	**3.9**	**30.52**	**0.944**
$\tau = 100$	9.4	33.08	0.978	**3.9**	30.16	0.943

It is worth to mention that the performance of shadow removal is highly affected by the predicted shadow mask obtained via [43]. We show some failure cases caused by the false detected shadow masks in Fig. 4. If a shadow is not detected, it cannot be removed, as shown in the top of Fig. 4. If a non-shadow region is mis-detected as shadow, it may become brighter after shadow removal, as shown in the bottom of Fig. 4. If we use the ground-truth shadow masks as input for testing, the performance of our model can be further improved. We conduct this experiment and find that RMSE, PSNR and SSIM of the predicted results from our proposed method can reach 8.6, 34.01, and 0.979, respectively, when evaluated on the shadow region of ISTD.

4.4. Comparison with the state-of-the-arts

In this subsection, we compare our proposed weakly-supervised method with several state-of-the-art methods on ISTD. In addition, we train our method on paired

Table 5. Quantitative comparison results of the proposed G2R-ShadowNet with the state-of-the-art methods. 'Non-Shadow Region' indicates that RMSE is computed on the non-shadow region of the testing images. 'RMSE*' indicates that RMSE is calculated by averaging the RMSE of all pixels in the shadow regions over the whole testing set. The results of these methods are either obtained from their original publications or produced by us using their official codes (marked with '*').

Method	Training Data	Shadow Region				Non-Shadow Region			All		
		RMSE*	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM
Yang *et al*. [37]	-	24.7	23.2	21.57	0.878	14.2	22.25	0.782	15.9	20.26	0.706
Gong and Cosker [5]	-	13.3	13.0	30.53	0.972	**2.6**	**36.63**	**0.982**	4.3	28.96	0.943
Guo *et al*. [8]	Shd.Free+Shd.Mask	22.0	20.1	26.89	0.960	3.1	35.48	0.975	6.1	25.51	0.924
ST-CGAN [32]	Shd.Free+Shd.Mask	13.4	12.0	31.70	0.979	7.9	26.39	0.956	8.6	24.75	0.927
SP+M-Net [16]	Shd.Free+Shd.Mask	7.9	8.1	35.08	0.984	2.8	36.38	0.979	**3.6**	31.89	0.953
G2R-ShadowNet *Sup.*	Shd.Free+Shd.Mask	**7.3**	**7.9**	**36.12**	**0.988**	2.9	35.21	0.977	**3.6**	**31.93**	**0.957**
Mask-ShadowGAN* [32]	Shd.Free (Unpaired)	9.9	10.8	32.19	0.984	3.8	33.44	0.974	4.8	28.81	0.946
LG-ShadowNet [22]	Shd.Free (Unpaired)	9.7	9.9	32.44	0.982	3.4	33.68	0.971	4.4	29.20	0.945
Le and Samaras [17]	Shd.Mask	9.7	10.4	33.09	0.983	2.9	35.26	0.977	4.0	30.12	0.950
G2R-ShadowNet	Shd.Mask	8.8	8.9	33.58	0.979	2.9	35.52	0.976	3.9	30.52	0.944

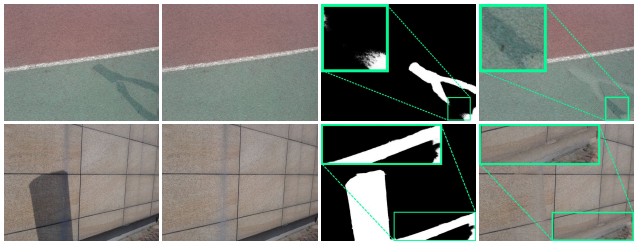

Figure 4. Failure cases on the ISTD dataset. The detected masks are generated by the shadow detector [43].

data by skipping the generation and training directly with shadow/non-shadow images from ISTD, and denote this model as G2R-ShadowNet *Sup.*.

The methods that we are compared with include Gong and Cosker [5], Guo *et al*. [8], Yang *et al*. [37], ST-CGAN [32], Mask-ShadowGAN [10], SP+M-Net [16], LG-ShadowNet [22], and Le and Samaras [17] on ISTD. Among them, Guo *et al*. [8], Yang *et al*. [37], and Gong and Cosker [5] use the pre-calculated image priors for shadow removal. ST-CGAN [32] and SP+M-Net [16] leverage paired shadow and shadow-free images, as well as shadow masks to train their models. Mask-ShadowGAN [10] and LG-ShadowNet [22] require unpaired shadow and shadow-free images for training. Le and Samaras [17] needs shadow images and shadow masks to train their network and our method use the same type of data as Le and Samaras [17].

Quantitative results are shown in Table 5. From the first block, we observe that our method outperforms the methods using the pre-calculated image priors except for the non-shadow region. Note that Gong and Cosker [5] use an interactive method, which requires user input, to define the shadow and non-shadow regions in testing, while we only use the one automatically generated by [43]. The methods in the second block share the same type of training data, including both shadow-free images and the shadow masks. Our weakly-supervised method achieves competitive performance with these methods but using less training data: we do not use shadow-free images. We also observe that our method trained on paired data, i.e., G2R-ShadowNet *Sup.*, outperforms all the methods in the shadow region and the whole image. In the third block, both Mask-ShadowGAN and LG-ShadowNet train their shadow removal models using unpaired shadow and shadow-free images. We can see that our method outperforms these two methods.

The comparison to Le and Samaras [17] is more fair since it is the only previous work that trains the model without using shadow-free images, which is also main goal of our method. From the last block, we can see that our method outperforms [17] on most metrics except that two of the SSIM values are slightly below [17].

Figure 5 shows the qualitative results of our method and the other state-of-the-art methods on four challenging samples drawn from the testing set of ISTD. Compared with other methods, our method can produce results with less artefacts. Moreover, the colour in the shadow region is more consistent with the surrounding area using our method, while the patch-based method [17] tends to produce over-lightened colour in the non-shadow region (column 5), making them easy to distinguish even after the shadow removal.

4.5. Generalisation ability

Finally, we show the generalisation ability of the proposed G2R-ShadowNet by comparing it with Mask-ShadowGAN [10], SP+M-Net [16], LG-ShadowNet [22], and Le and Samaras [17]. Here all methods are trained on ISTD and tested on the video dataset without additional training or fine-tuning. The quantitative results are reported in Table 6.

We observe that our method outperforms Mask-

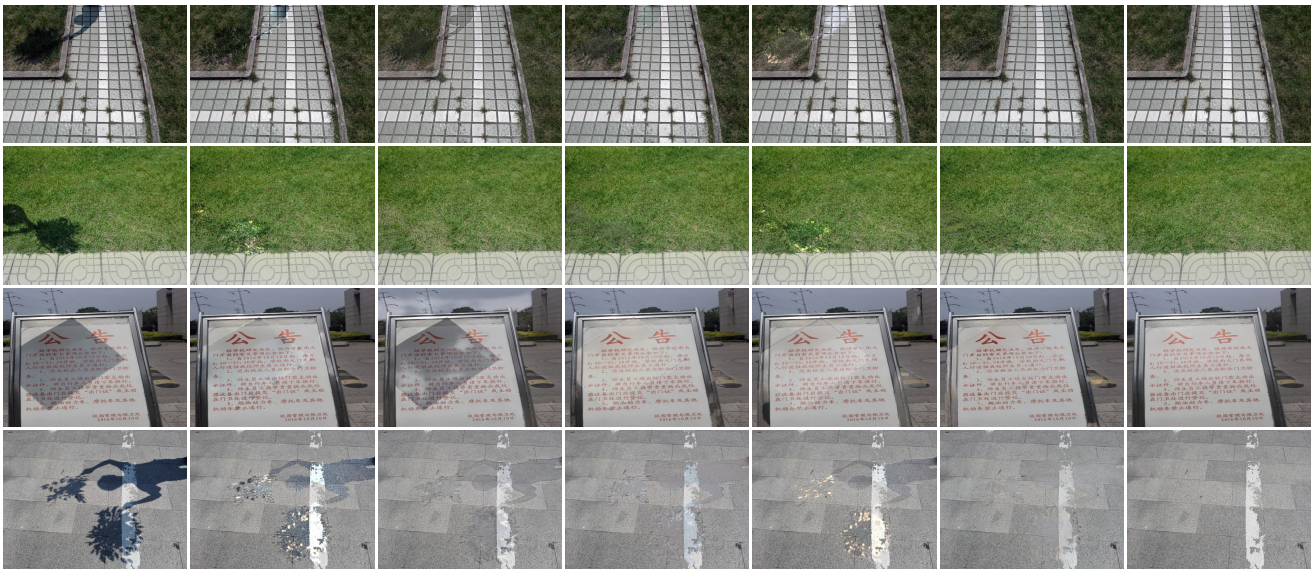

Figure 5. Visualisation comparisons on four challenging samples from the testing set of the ISTD dataset.

Table 6. Quantitative comparison of the generalisation ability of the proposed G2R-ShadowNet and the state-of-the-art methods on the video shadow removal dataset. Note that we compute the metrics only in the moving-shadow region. 'RMSE†' is the RMSE computed by using the moving-shadow mask with a threshold of 40, while other metrics are computed using a threshold of 80. '-' in the first two rows means the results are not publicly available.

Method	RMSE	RMSE†	PSNR	SSIM
SP+M-Net [16]	-	22.2	-	-
Le and Samaras [17]	-	20.9	-	-
Mask-ShadowGAN* [10]	22.7	19.6	20.38	**0.887**
LG-ShadowNet* [22]	<u>22.0</u>	**18.3**	<u>20.68</u>	0.880
G2R-ShadowNet (Ours)	**21.8**	<u>18.8</u>	**21.07**	<u>0.882</u>

ShadowGAN [10] and LG-ShadowNet [22] significantly on RMSE and PSNR metrics, indicating that our method has better generalisation ability on other unseen environments. We also fine-tune our model on each testing video for additional 1 epoch and it further improves the performance gains on RMSE by about 14% (from 21.8 to 18.7).

We also show visualisation comparison results with Mask-ShadowGAN [10] and LG-ShadowNet [22] on two samples from the video dataset in Fig. 6. The shadow regions in our results look lighter with less artefacts than others for images of either close (top) or distant shots (bottom).

Figure 6. Visualisation comparisons on two sample images from the Video Shadow Removal dataset [17] with the Mask-ShadowGAN [10] and LG-ShadowNet [22].

5. Conclusion

To conclude, we proposed a novel G2R-ShadowNet for weakly-supervised shadow removal which is trained without using shadow-free images. The training of the network consists of shadow generation, shadow removal and refinement, which correspond to three sub-nets in G2R-ShadowNet, respectively, and they are jointly trained in an end-to-end fashion. Shadow generation is a prerequisite which stylises non-shadow regions to be shadow ones and constructs paired training set for shadow removal. Extensive experiments showed the effectiveness of our proposed G2R-ShadowNet and verified that our method outperforms the best weakly-supervised method on the adjusted ISTD dataset and the Video Shadow Removal dataset. It also achieved competitive performances against the other state-of-the-arts that use more training data, such as paired or unpaired shadow-free images, than our method.

Acknowledgements. This work was partially supported by the Research and Development Program of Beijing Municipal Education Commission (KJZD20191000402) and by the National Nature Science Foundation of China (51827813, 61472029, 61672376, U1803264).

References

[1] Tim Brooks and Jonathan T Barron. Learning to synthesize motion blur. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6840–6848, 2019. 2

[2] Rita Cucchiara, Costantino Grana, Massimo Piccardi, and Andrea Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(10):1337–1342, 2003. 1

[3] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In *Int. Conf. Comput. Vis.*, 2019. 2

[4] Graham D Finlayson, Steven D Hordley, Cheng Lu, and Mark S Drew. On the removal of shadows from images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(1):59–68, 2005. 1, 2

[5] Han Gong and Darren Cosker. Interactive shadow removal and ground truth for variable scene categories. In *Brit. Mach. Vis. Conf.*, 2014. 7, 8

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Adv. Neural Inform. Process. Syst.*, 2014. 2

[7] Maciej Gryka, Michael Terry, and Gabriel J Brostow. Learning to remove soft shadows. *ACM Trans. Graph.*, 34(5):153, 2015. 2

[8] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2956–2967, 2012. 1, 2, 7

[9] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 1, 2

[10] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *Int. Conf. Comput. Vis.*, 2019. 1, 2, 3, 5, 7, 8

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 3

[12] Cláudio Rosito Jung. Efficient background subtraction and shadow removal for monochromatic video sequences. *IEEE Transactions on Multimedia*, 11(3):571–577, 2009. 1

[13] Salman H Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Automatic shadow detection and removal from a single image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(3):431–446, 2015. 1

[14] Hieu Le, Bento Goncalves, Dimitris Samaras, and Heather Lynch. Weakly labeling the antarctic: The penguin colony case. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 18–25, 2019. 1

[15] Hieu Le, Vu Nguyen, Chen-Ping Yu, and Dimitris Samaras. Geodesic distance histogram feature for video segmentation. In *ACCV*, pages 275–290. Springer, 2016. 1

[16] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *Int. Conf. Comput. Vis.*, 2019. 1, 2, 4, 5, 7, 8

[17] Hieu Le and Dimitris Samaras. From shadow segmentation to shadow removal. In *Eur. Conf. Comput. Vis.*, 2020. 1, 2, 5, 7, 8

[18] 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 *Eur. Conf. Comput. Vis.*, 2018. 1, 2

[19] Hieu Le, Chen-Ping Yu, Gregory Zelinsky, and Dimitris Samaras. Co-localization with category-consistent features and geodesic distance propagation. In *Int. Conf. Comput. Vis. Worksh.*, pages 1103–1112, 2017. 1

[20] Yun-Hsuan Lin, Wen-Chin Chen, and Yung-Yu Chuang. Bedsr-net: A deep shadow removal network from a single document image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12905–12914, 2020. 1, 2

[21] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8139–8148, 2020. 2

[22] Zhihao Liu, Hui Yin, Yang Mi, Mengyang Pu, and Song Wang. Shadow removal by a lightness-guided network with training on unpaired data. *IEEE Trans. Image Process.*, 30:1853–1865, 2021. 1, 2, 5, 7, 8

[23] Thomas Müller and Bastian Erdnüeß. Brightness correction and shadow removal for video change detection with uavs. In *Autonomous Systems: Sensors, Processing, and Security for Vehicles and Infrastructure 2019*, volume 11009, page 1100906. International Society for Optics and Photonics, 2019. 1

[24] Sohail Nadimi and Bir Bhanu. Physical models for moving shadow and object detection in video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(8):1079–1087, 2004. 1

[25] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1, 2

[26] Yael Shor and Dani Lischinski. The shadow meets the mask: Pyramid-based shadow removal. *Comput. Graph. Forum*, 27:577–586, 04 2008. 2

[27] Nan Su, Ye Zhang, Shu Tian, Yiming Yan, and Xinyuan Miao. Shadow detection and removal for occluded object information recovery in urban high-resolution panchromatic satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(6):2568–2582, 2016. 1

[28] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 4

[29] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 3

[30] Tomas F Yago Vicente, Minh Hoai, and Dimitris Samaras. Leave-one-out kernel optimization for shadow detec-

tion and removal. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(3):682–695, 2017. 2

[31] 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 *Eur. Conf. Comput. Vis.*, pages 816–832. Springer, 2016. 5

[32] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1, 2, 4, 5, 7

[33] Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1880–1889, 2020. 2

[34] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3771–3779, 2019. 2

[35] Chunxia Xiao, Ruiyun She, Donglin Xiao, and Kwan-Liu Ma. Fast shadow removal using adaptive multi-scale illumination transfer. In *Comput. Graph. Forum*, 2013. 2

[36] Jun Xu, Yuan Huang, Ming-Ming Cheng, Li Liu, Fan Zhu, Zhou Xu, and Ling Shao. Noisy-as-clean: Learning self-supervised denoising from corrupted image. *IEEE Trans. Image Process.*, 29:9316–9329, 2020. 2

[37] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. *IEEE Trans. Image Process.*, 21(10):4361–4368, 2012. 2, 7

[38] Ling Zhang, Qing Zhang, and Chunxia Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Trans. Image Process.*, 24(11):4623–4636, 2015. 1, 2

[39] Shuyang Zhang, Runze Liang, and Miao Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5(1):105–115, 2019. 2

[40] Wuming Zhang, Xi Zhao, Jean-Marie Morvan, and Liming Chen. Improving shadow suppression for illumination robust face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(3):611–624, 2018. 1

[41] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. Distraction-aware shadow detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5167–5176, 2019. 2

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Int. Conf. Comput. Vis.*, 2017. 2, 4

[43] 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 *Eur. Conf. Comput. Vis.*, pages 121–136, 2018. 2, 5, 6, 7