


Depth-Aware Shadow Removal

Yanping Fu¹ , Zhenyu Gai¹, Haifeng Zhao¹, Shaojie Zhang¹, Ying Shan², Yang Wu² and Jin Tang¹ †

¹Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University

²ARC Lab, Tencent PC

ypfu@ahu.edu.cn, e20301305@stu.ahu.edu.cn, senith@163.com, zhangshaojie@ahu.edu.cn, {yingsshan, dylanywu}@tencent.com, tangjin@ahu.edu.cn

Abstract

Shadow removal from a single image is an ill-posed problem because shadow generation is affected by the complex interactions of geometry, albedo, and illumination. Most recent deep learning-based methods try to directly estimate the mapping between the non-shadow and shadow image pairs to predict the shadow-free image. However, they are not very effective for shadow images with complex shadows or messy backgrounds. In this paper, we propose a novel end-to-end depth-aware shadow removal method without using depth images, which estimates depth information from RGB images and leverages the depth feature as guidance to enhance shadow removal and refinement. The proposed framework consists of three components, including depth prediction, shadow removal, and boundary refinement. First, the depth prediction module is used to predict the corresponding depth map of the input shadow image. Then, we propose a new generative adversarial network (GAN) method integrated with depth information to remove shadows in the RGB image. Finally, we propose an effective boundary refinement framework to alleviate the artifact around boundaries after shadow removal by depth cues. We conduct experiments on several public datasets and real-world shadow images. The experimental results demonstrate the efficiency of the proposed method and superior performance against state-of-the-art methods.

CCS Concepts

• *Computing methodologies* → *Image processing; Computational photography;*

1. Introduction

Shadow removal is a fundamental and challenging task in computer vision. Shadow is a very common natural phenomenon in daily life, which is caused by light being partially or completely blocked. Therefore, we inevitably obtain a number of shadow images and shadow videos when we use cameras or smartphones. The shadow in these images and videos will have a certain impact on computer vision tasks, such as visual odometry [MAT17, EKC17], object detection and tracking [XRG*17, CGPP03, NB04, HFY20], relighting [WSL*20, YME*20], and object recognition [NB04, HFY20], etc. The aim of shadow removal is to restore the illumination, color, and texture in the shadow regions. Therefore, the challenge of shadow removal is how to restore the original detail and keep consistency in the shadow region after shadow removal.

Previous shadow removal methods are generally based on physical models and adopt the prior information such as gradient, illumination, and regions to remove shadows. However, these hand-crafted shadow removal methods can not achieve satisfactory results for shadow images with complex shadows or cluttered backgrounds. Xiao *et al.* [XTT14] first introduced the depth infor-

mation into the shadow removal task, which assumed that pixels with similar normal and close spatial locations should have similar color and illumination. This method can generate satisfactory shadow-free images for complex scenes. However, the image obtained in advance usually has no corresponding depth map, which is difficult to be extended to practical applications. In recent years, the emergence of deep learning has greatly improved the performance of shadow removal. Auto-Exposure [FZG*21] first estimated multi-exposure images for the shadow image. Then they fused multi-exposure images to predict the shadow-free image. G2R-ShadowNet [LYW*21] used a weakly supervised shadow generation network for shadow removal by using a set of shadow images and corresponding shadow masks without shadow-free image supervision. DC-ShadowNet [JST21] proposed an unsupervised domain-classifier-guided shadow removal network, which integrated the domain classifier into GAN to predict the shadow-free image. These deep learning-based methods are very effective for simple scenes and shadows. However, for complex scenes and shadows, these methods are hard to obtain satisfactory shadow removal results, as shown in Figure 1.

To solve the aforementioned issues, we propose a novel end-to-end depth-aware shadow removal method, which can effectively handle shadow images with complex shadows or backgrounds

† Corresponding author

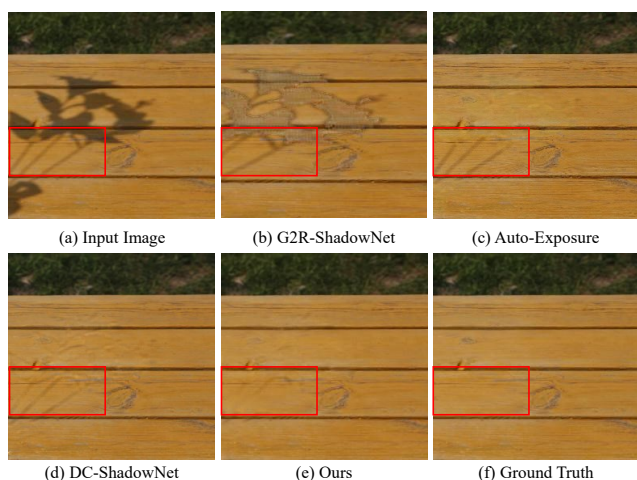


Figure 1: Shadow removal results of state-of-the-art methods, including Mask-ShadowGAN [HJFH19], Auto-Exposure [FZG*21], DC-ShadowNet [JST21] and the proposed method.

based on the guidance of predicted depth information. The pipeline of the proposed method contains three modules: depth prediction, shadow removal, and boundary refinement. First, we introduce a depth-aware network to predict the depth information for the shadow image. Second, we use the predicted depth image as structural guidance and propose a multi-modality shadow removal architecture based on the generative adversarial network (GAN). We extract features from the color image and the predicted depth image respectively, and design an effective cross-modal feature fusion module for shadow removal. Finally, we design a boundary refinement module using predicted depth cues to alleviate the artifacts around shadow boundaries. We demonstrate the effectiveness of the proposed method on the public datasets (ISTD [WLY18], SRD [QTH*17]) and several real-world scenes. The experimental results show that the proposed method achieves superior performance to state-of-the-art methods.

The main contributions of this work are listed as follows:

- We propose a novel end-to-end depth-aware shadow removal method without requiring the existence of depth images, which can effectively deal with shadow removal for complex scenes.
- We propose a boundary refinement strategy using depth cues to mitigate the artifacts around shadow boundaries caused by shadow removal.
- The experimental results demonstrate that our method achieves leading shadow removal performance in qualitative and quantitative evaluation.

2. Related Work

Due to shadows widely existing in images and videos, shadow removal methods have been extensively studied in recent decades. In this section, we revisit some recent approaches that are closely related to the proposed method.

Hand-crafted shadow removal. Early shadow removal methods

are mostly based on physical models and the prior information, such as illumination [SL08, ZZ15a, FWZ*20, ZZ15b], gradient [FHL05, FDL09] and color transfer [VHS17, WTBS07]. Some other methods adopt user interaction for shadow removal tasks. Gong *et al.* [GC17] proposed a shadow removal method using user-defined flexible strokes covering the shadow and non-shadow pixels. Murali *et al.* [MGK21] presented an interactive technique for shadow removal from images, which also needed user input in the form of rough strokes on the shadow region and its corresponding non-shadow region. However, shadows are highly coupled with geometry, albedo, and illumination. These physical-based methods only model the intrinsic image priors without considering geometric information. Therefore, it is difficult to achieve promising results for some complex shadows and scenes. Xiao *et al.* [XTT14] first introduced depth information to assist shadow removal, which was based on the assumption that pixels with similar normals and locations should have similar colors. However, this method needs to obtain the corresponding depth image with the shadow image in advance, which greatly limits its practical application.

Learning-based shadow removal. In recent years, more and more methods [ZLZX20, VRVGT21, LYM*21, CLZX21] have begun to use deep learning to improve the performance of shadow removal owing to the powerful representation ability of CNN. Zhu *et al.* [ZXF*22] proposed a new shadow illumination model considering the spatially-variant property for shadow removal. They reformulated shadow removal as a variational optimization problem. Le *et al.* [LS20] proposed a patch-based deep learning model to remove the shadow from images, which can be trained using shadow and non-shadow patches cropped from shadow images. Wang *et al.* [WLY18] proposed a stacked conditional generative adversarial network (ST-CGAN) to jointly learn shadow detection and shadow removal. They also proposed a public dataset ISTD to train the network. BEDSR-Net [LCC20] used the special attributes of the document image to recover the shadow areas by attention mechanism. DSC [HFZ*19] used direction-aware spatial context (DSC) for detecting and removing shadows. Mask-ShadowGAN [HJFH19] proposed a mask-guided generative adversarial network, which learned to produce a shadow mask from the input and took the mask to guide the shadow generation and removal. Liu *et al.* [LYW*21] proposed a network named G2R-ShadowNet, which leveraged shadow generation for weakly-supervised shadow removal using a set of shadow images and their corresponding shadow masks. These learning-based methods have achieved promising performances in shadow removal.

Depth-aware methods. As complementary information to RGB images, some works have begun to pay attention to use depth map as complementary information to improve the performance of computer vision tasks. However, most captured color images do not have corresponding depth images. Therefore, many vision tasks cannot achieve satisfactory results. So the depth estimation from color images based on deep learning has become an alternative solution for vision tasks in recent years and has been widely studied. Hu *et al.* [HZW*21] designed an end-to-end deep neural network to learn the depth-guided non-local features and produce a rain-free output image. Qian *et al.* [QYL*21] proposed a deep depth-aware long-term tracker and achieved state-of-the-art tracking performance. Zhang *et al.* [ZZJ*21] proposed a depth predic-

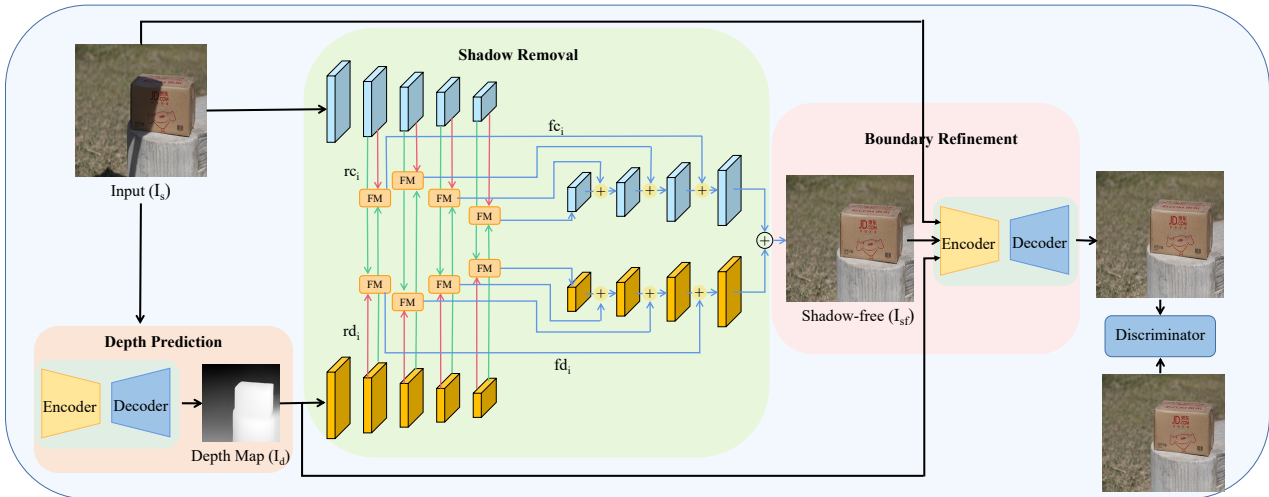


Figure 2: The overview of the proposed network architecture.

tion network to estimate the depth information and leveraged the depth feature to enhance the saliency detection performance. Inspired by these works, we propose a novel depth-aware shadow removal method and use the depth feature as guidance to improve the shadow removal performance.

3. Methodology

In this section, we will illustrate our proposed network architecture in detail. Intuitively, the depth information in the shadow regions and non-shadow regions will not be affected by shadow. Therefore, we propose an end-to-end depth-aware shadow removal method, including three modules: depth prediction, shadow removal, and boundary refinement. The depth prediction network is pre-trained, and the shadow removal and boundary refinement networks are trained in an end-to-end manner. Figure 2 shows the network architecture of the proposed method.

3.1. Depth Prediction Module

Most traditional shadow removal methods adopt intrinsic priors of the image such as gradient and illumination to explore the physical properties. Although these methods are promising for simple scene shadow images, they are limited to some simple shadow images. The main reason is that these methods intend to ignore the geometry prior, although it is closely related to shadow generation. Because the depth image needs an extra depth sensor, and most of the images acquired in advance have no corresponding depth images. Therefore, the traditional shadow removal methods rarely adopt depth prior. With the rise of deep learning, it is possible to obtain a predicted depth image from a single RGB image [RBK21, DC19, XZW*20]. In order to better handle shadow images, we propose a novel shadow removal network that uses the predicted depth maps as guidance.

For depth prediction, we introduce DPT-Net [RBK21] to predict the corresponding depth map to the input shadow image. DPT-Net is a dense prediction architecture based on a vision transformer

(ViT). They utilize the vision transformer substitute for convolutional networks as the basic backbone. The backbone of DPT-Net processes image-like feature representations at high resolution with a global receptive field at every stage, which promises to provide finer-grained and globally coherent predictions. However, due to the illumination degradation in the shadow regions, it may generate unsatisfactory results if we directly use this method in shadow image depth prediction. Therefore, we create a synthetic dataset to fine-tune the DPT-Net so that the network can adapt to the depth prediction for shadow images. The depth prediction results of shadow images are shown in Figure 3.

To fine-tune the depth prediction network, we use the depth estimation dataset NYU Depth v2 [SHKF12] to synthesize the depth prediction dataset for shadow images. NYU Depth v2 dataset consists of paired color images and depth images, which is unsuitable for our shadow image depth prediction. First, we randomly select a shadow mask from the shadow removal dataset ISTD [WLY18]. Then, inspired by G2R-ShadowNet [LYW*21], we use the shadow generation method to generate a shadow using the selected shadow mask onto the color image selected from the NYU Depth v2 dataset. Finally, we can obtain the dataset composed of triplets of shadow images, shadow masks, and ground truth depth images. Then we use this synthesis dataset to fine-tune the DPT-Net. After training, the parameters of the depth prediction network will be fixed in all the next experiments.

3.2. Shadow Removal Module

We use a generative adversarial network (GAN) to design the shadow removal module, which includes a generator and a discriminator. Furthermore, to make full use of the complementary information between depth and color features to improve the performance of shadow removal, we introduce both multi-scale feature fusion and cross-modal feature fusion strategies into our dual encoder-decoder architecture.

We adopt a symmetric encoder-decoder architecture to construct



Figure 3: The shadow images and the predicted depth maps.

the generator. The generator consists of two parallel backbone networks: the encoder-decoder of the depth image and the encoder-decoder of the shadow image. Both networks adopt the same encoder and decoder architectures. In the encoder stage, a cross-modal fusion strategy is used to capture inter-dependent complementary information between the depth and appearance features. In the decoder stage, multi-scale feature fusion is used to aggregate more detailed information to improve performance. At the end of the decoder, the cross-modal feature maps merged at the top of the two decoders are used to aggregate and output the predicted non-shadow images. Next, the non-shadow image is sent to the boundary refinement module to refine the shadow boundaries. Finally, the predicted non-shadow image is fed into the discriminator for identification. We use a discriminator and the ground truth shadow-free image to supervise and train the network. The pipeline of the shadow removal module is shown in Figure 2.

The input of the shadow image backbone network is the shadow image, and the input of the depth backbone network is the depth image predicted from the depth prediction module. For simplicity, we also define the output feature of each encoder component in the shadow image backbone network as rc_i and in the depth backbone network as rd_i . The output of the fusion module is fc_i or fd_i , where $i (i = 1, 2, 3, 4, 5)$ represents the index of convolution layer.

Fusion Module (FM). To make full use of the complementary information between the depth image feature and the shadow image feature, we introduce a cross-modal fusion strategy according to the attention mechanism proposed by [CCXH20], as shown in Figure 4.

Figure 4 shows the architecture of the fusion module. f_1 indicates the output rd_i or rc_i , and f_2 represents the opposite one. We first feed f_1 and f_2 into a spatial attention to highlight the feature response, and use the ReLU activation function to obtain the outputs f_3 and f_4 , respectively. Then we send f_4 into a self-attention model to capture a spatial weight for f_3 from a cross-modal perspective. This sub-module can make RGB and depth information features provide useful information to each other. These feature maps f_1 and f_2 are fed into two attention modules, which can generate a spatial weight for the other feature map and provide complementary information for cross-modal fusion. The output of the fusion module can be defined as:

$$f_{out} = SP(f_i) + f, \quad (1)$$

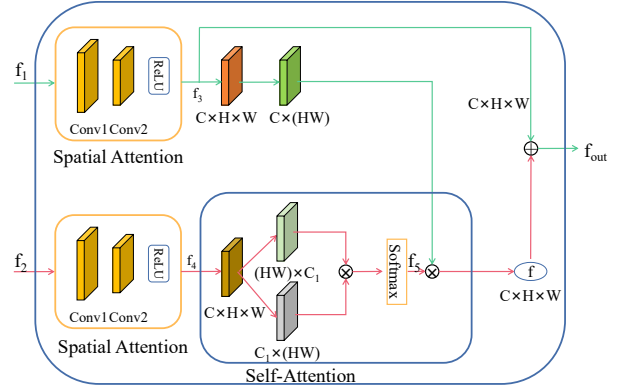


Figure 4: The fusion module (FM) architecture.

where SP means spatial attention operation. f_i represents rd_i or rc_i in turn.

For the shadow image feature rc_i , we feed it as f_1 and rd_i as f_2 into the fusion module to obtain the spatial weight from depth information. We exchange the input rc_i and rd_i , which can obtain the spatial weight from the shadow image feature for the depth feature.

Multi-scale Feature Fusion. The advantage of multi-scale fusion is that we can use different information contained at different levels to improve the results of our task. The low-level features can provide more detailed information, and the high-level features contain more semantic information. Therefore, we also use a multi-scale feature fusion strategy in the decoder stage. We use the output of the fusion model to skip connecting to the decoder, as shown in Figure 2.

3.3. Boundary Refinement Module

Depending on the different occlusion of the light source, the shadow usually forms umbra and penumbra regions. The umbra region is a hard shadow completely obscured by the occlusion object. The illumination in the umbra region is uniform, and most shadow removal algorithms can handle it very well. On the contrary, the penumbra region is formed by the diffraction of light at the boundary of occluding object. Therefore, the illumination in this region changes gradually, and the brightness in this region is nonuniform. Thus the penumbra region of the shadow, namely the soft shadow, is challenging to the existing shadow removal methods. Therefore, most state-of-the-art shadow removal methods will introduce artifacts or pseudo-color in these penumbra regions. In order to effectively alleviate the artifacts in the shadow removal, we propose a new boundary refinement network using depth cues to deal with this issue, as shown in Figure 5.

We design an encoder-decoder network that uses depth cues as a guide to optimizing shadow boundaries to make full use of depth information. The input to the network is the predicted shadow-free image (I_f), the original shadow image (I_s), and the predicted depth map (I_d). In order to better deal with the artifacts of shadow boundaries, we refer to the method of producing soft shadow masks

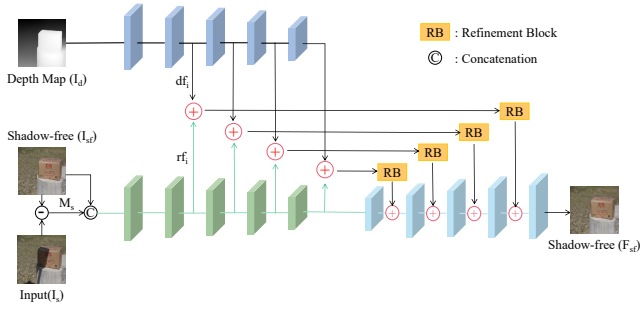


Figure 5: The network architecture of boundary refinement module.

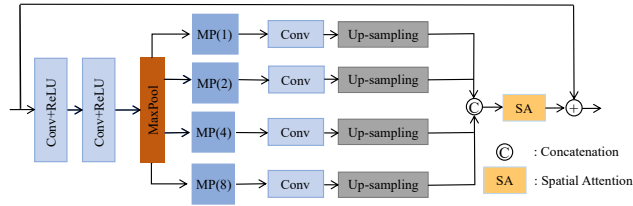


Figure 6: The network architecture of refinement block.

mentioned in the DC-ShadowNet [JST21] method. We generate a single-channel shadow mask (M_s) by computing the difference between the predicted shadow-free image (I_f) and the shadow image (I_s). Then, we concatenate the single-channel shadow mask with the predicted shadow-free image (I_f) as the input to the encoder branch. To use the depth map as guidance to refine the boundary of shadows, we feed the predicted depth map (I_d) into a similar encoder branch. In order to make full use of cross-modal features, we first combine two model features (df_i and rf_i) from each encoder layer $i = \{2, 3, 4, 5\}$. Then we feed the cross-modal feature to a refinement block (RB) to generate multi-scale cross-modal attention for shadow boundaries. Finally, we add multi-scale cross-modal attention to the decoder to refine the shadow boundaries.

Inspired by the method of [QCHX19], we design the refinement block module to capture multi-scale cross-modal attention for shadow boundaries, as shown in Figure 6. Since our main purpose is to optimize the shadow boundary, we use max-pooling (MP) to highlight the feature of the boundary. We also introduce different step sizes for max-pooling to change the size of feature maps and receptive fields, which can effectively extract multi-scale features to prevent information loss. Then, we upsample multi-scale features to the original size and concatenate them. In addition to this, we utilize a spatial attention mechanism to refine feature maps by spatially exploiting weights. Finally, we obtain the final output of the refinement block by a skip connection, which is effective for preventing image blur.

3.4. Loss Function

We use GAN to design the shadow removal architecture. The generator is composed of a multi-modality encoder-decoder network, and the discriminator D is constructed with five-layer convolutions. By distinguishing the difference between the predicted non-shadow

image and the ground truth shadow-free image, the discriminator is optimized to recognize the predicted image. Thus, the generator is promoted to generate a more realistic shadow-free image. We use L_1 loss for generator G to calculate the difference between the predicted non-shadow image and the ground truth to optimize the generator. The discriminator is mainly used to distinguish the difference between the predicted non-shadow image by the generator and the ground truth. Therefore, we use the binary cross-entropy loss function to optimize the discriminator. The objective functions of training the shadow generator G and discriminator D are defined as:

$$L_G = \lambda_1 |y - G(x, E(x))| + \lambda_2 \log D(G(x, E(x)), y) \quad (2)$$

$$L_D = \lambda_3 \log D(G(x, E(x)), y) + \lambda_4 \log (1 - D(G(x, E(x)), y)) \quad (3)$$

where x represents the shadow image, $E(x)$ represents the depth image predicted from depth prediction module, y is the ground truth shadow-free image. In all experiments, we set $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ to 5, 0.1, 0.1, 0.1, respectively.

4. Experiments

4.1. Implementation details

The proposed network is implemented with PyTorch on a PC with 8 NVIDIA GeForce GTX 1080Ti GPU. In our experiments, we use the Adam optimizer with the batch size 8 and 1500 epochs to train the proposed network. The first momentum value and the second momentum value are set to 0.5 and 0.999, respectively. The initial learning rate is set to 2×10^{-4} . We apply random cropping and flipping to the shadow image for data enhancement to avoid overfitting problems. The random cropping is achieved by first scaling each image to 286×286 , and then randomly cropping a 256×256 area from the scaled result. To ensure a fair comparison, we use the same input image size for all methods.

4.2. Datasets and evaluation metrics

We train and evaluate the proposed method on two public datasets: ISTD [WLY18] and SRD [QTH*17].

ISTD. The ISTD dataset is proposed for shadow detection and shadow removal, which is collected under different lighting conditions with varying shapes of shadow. The ISTD dataset is composed of image triplets, including shadow image, shadow mask, and shadow-free image. The training set has 1870 image triplets from 135 various scenes, and the test set has 540 image triplets from 45 different scenes.

SRD. The SRD dataset consists of shadow and shadow-free image pairs. We use 2680 shadow from SRD image pairs for training and 408 shadow image pairs for testing.

Evaluation Metrics. For qualitative and quantitative analysis, we use root mean square error (RMSE), structural similarity (SSIM), and peak signal-to-noise ratio (PSNR) as the evaluation metrics on ISTD and SRD datasets. The RMSE calculates the root mean square error between the ground truth shadow-free image and the

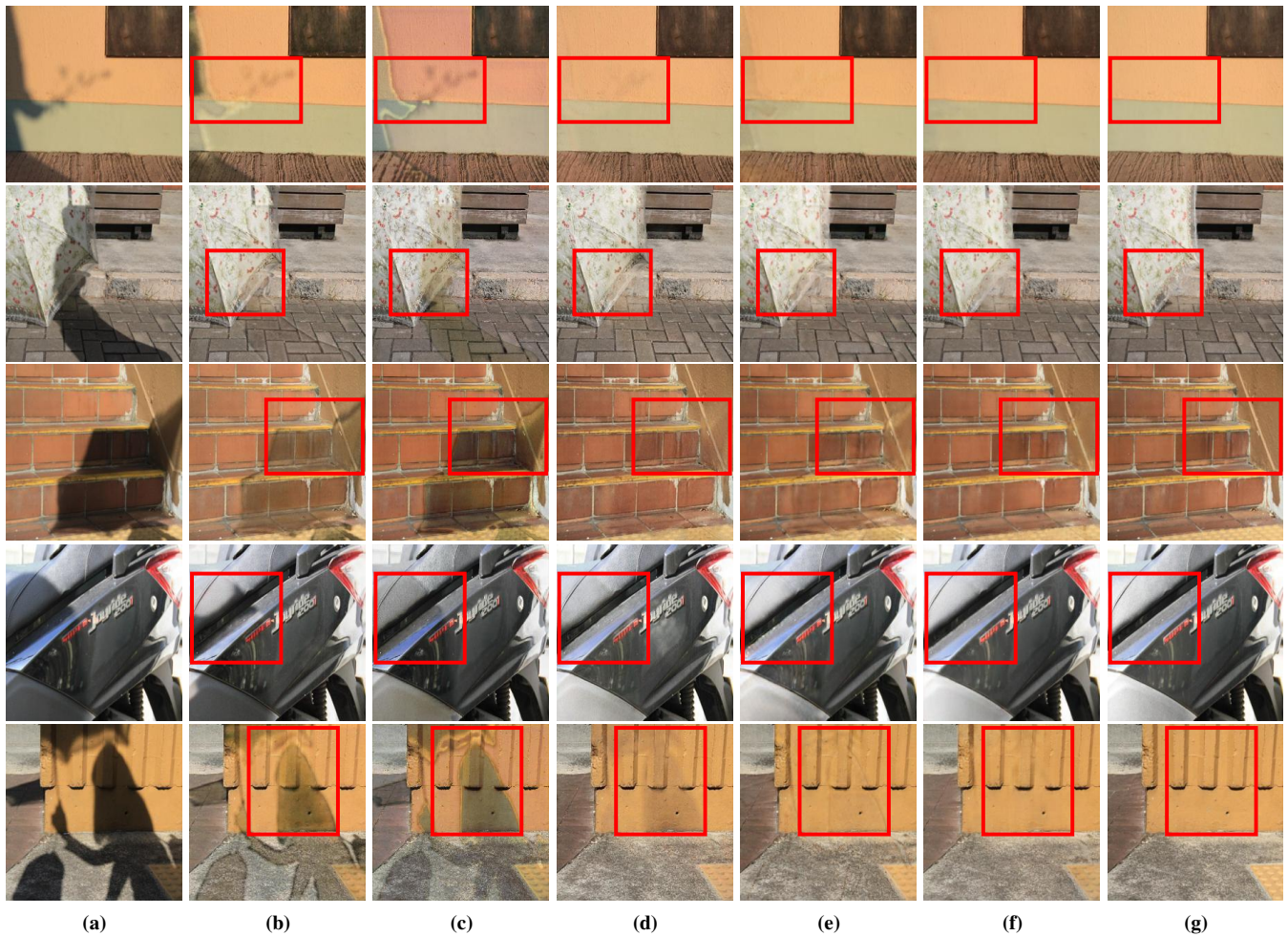


Figure 7: Visual comparison results of shadow removal on SRD dataset. (a) Input. (b) Mask-ShadowGAN [HJFH19]. (c) Auto-Exposure [FZG*21]. (d) DHAN [CPS20]. (e) DC-Shadow [JST21]. (f) Ours. (g) Ground Truth.

predicted non-shadow image in the LAB color space. Generally speaking, the smaller value of RMSE means the better performance of the shadow removal method. We also calculate the SSIM and PSNR metrics in RGB space to further demonstrate the effectiveness of the proposed method.

4.3. Shadow removal evaluation on SRD dataset

We first compare the shadow removal results of the proposed method on the SRD dataset with state-of-the-art methods, including the methods of Gong *et al.* [GC14], DSC [HFZ*19], DeShadowNet [QTH*17], DHAN [CPS20], Auto-Exposure [FZG*21], DC-ShadowNet [JST21], and Zhu *et al.* [ZXF*22]. To demonstrate the effectiveness of the proposed methods, we report three metrics (SSIM, PSNR, and RMSE) values for each method in the shadow region, non-shadow region, and the whole image (All), respectively. For a fair comparison, all metric values are provided by the authors and their reports in their manuscripts. As can be seen from Table 1, the proposed method exceeds state-of-the-art

methods. The proposed method outperforms existing methods on both metrics, PSNR and RMSE, whether in shadow, non-shadow, or all image regions. For the SSIM metric, the method is comparable to other methods. The results show that our method has the best shadow removal performance in both shadowed and non-shadowed regions, resulting in the lowest RMSE for the entire image.

Figure 7 shows the visual comparison results of the proposed method and other state-of-the-art methods for shadow removal on the SRD dataset. We can see that the results of Mask-ShadowGAN [HJFH19], Auto-Exposure [FZG*21] and DHAN [CPS20] methods are obvious artifacts in the shadow regions. The results of DC-Shadow [JST21] also produce some slight inconsistencies in the non-shadow regions, as shown in Figure 7(e). However, the proposed method can recover high-fidelity backgrounds in shadow regions and obtain consistent shadow removal results on boundaries, as shown in Figure 7(f).

Table 1: Shadow removal results of our method compared to state-of-the-art shadow removal methods on the SRD dataset. The best and the second best values are marked with **bold** and underline, respectively.

Method	Shadow Region			Non-Shadow Region			All		
	SSIM \uparrow	PSNR \uparrow	RMSE \downarrow	SSIM \uparrow	PSNR \uparrow	RMSE \downarrow	SSIM \uparrow	PSNR \uparrow	RMSE \downarrow
Gong <i>et al.</i> [GC14]	-	-	25.43	-	-	6.91	-	-	12.35
DSC [HFZ*19]	0.904	26.42	10.89	0.773	24.67	4.99	0.655	21.52	6.23
DeShadowNet [JST21]	0.947	31.97	10.81	-	-	4.85	-	-	6.10
DHAN [CPS20]	0.978	33.67	8.98	<u>0.979</u>	34.79	4.80	0.949	30.51	5.67
Auto-Exposure [FZG*21]	0.966	32.26	8.55	0.945	30.59	5.74	0.893	27.74	6.50
DC-ShadowNet [GC14]	-	-	7.70	-	-	<u>3.39</u>	-	-	<u>4.66</u>
Zhu <i>et al.</i> [ZXF*22]	<u>0.979</u>	<u>34.94</u>	<u>7.44</u>	0.981	<u>35.85</u>	3.74	0.952	<u>31.72</u>	4.79
Ours	0.982	42.48	4.59	0.960	38.69	3.38	<u>0.934</u>	36.19	3.69

Table 2: Shadow removal results of our method compared to state-of-the-art shadow removal methods on the ISTD dataset.

Method	Shadow Region			Non-Shadow Region			All		
	SSIM \uparrow	PSNR \uparrow	RMSE \downarrow	SSIM \uparrow	PSNR \uparrow	RMSE \downarrow	SSIM \uparrow	PSNR \uparrow	RMSE \downarrow
Guo [GDH12]	0.960	26.89	18.65	0.975	35.48	7.76	0.924	25.51	9.26
ST-CGAN [WLY18]	0.979	31.70	10.33	0.956	26.39	6.93	0.927	24.75	7.47
SP-M-Net [LS19]	0.984	35.08	10.30	0.979	36.38	7.47	0.953	31.89	7.79
Mask-ShadowGAN [HJFH19]	0.984	32.19	12.67	0.974	33.44	6.68	0.946	28.81	7.41
LG-ShadowNet [LYM*21]	0.982	32.44	11.32	0.971	33.68	8.05	0.945	29.20	8.35
DSC [HFZ*19]	0.984	34.71	8.70	0.970	31.27	5.10	0.944	29.08	5.61
Le and Samaras [LS20]	0.983	33.09	11.82	<u>0.977</u>	35.26	7.53	0.950	30.12	7.94
G2R-ShadowNet [LYW*21]	0.988	36.12	<u>6.75</u>	<u>0.977</u>	35.21	4.78	0.957	<u>31.93</u>	5.15
Auto-Exposure [FZG*21]	0.959	34.66	7.77	0.904	26.22	5.56	0.897	25.81	5.92
Zhu <i>et al.</i> [ZXF*22]	<u>0.986</u>	<u>36.95</u>	8.29	<u>0.977</u>	31.54	<u>4.55</u>	<u>0.959</u>	29.85	<u>5.09</u>
Ours	0.981	41.59	4.92	0.976	<u>36.12</u>	3.91	0.963	34.34	4.15

4.4. Shadow removal evaluation on ISTD dataset

In this section, we compare the proposed method with several state-of-the-art shadow removal methods on the ISTD dataset, and the quantitative results are shown in Table 2. We use pretrained models of these methods to obtain evaluation metrics or author reports in their manuscripts. From Table 2, it can be seen that the proposed method has the lowest RMSE in the shadow area, non-shadow area and the whole image. This means that the results after shadow removal in our method are closer to the ground truth, that is, the proposed method has the best performance.

4.5. Real-scene dataset

To demonstrate the effectiveness of the proposed method, we also conduct experiments on real-scene datasets and compare the results with state-of-the-art shadow removal methods. For quantitative analysis, we fixed the camera position to take an image pair: shadow image and shadow-free image. First, we put an object under sunlight to make a shadow and take a shadow image. Then we take the object away to capture a shadow-free image.

Figure 8 shows the visual comparison results with state-of-the-art methods on real-scene datasets, which contain complex backgrounds and soft shadows. We can see that the method of Mask-ShadowGAN [HJFH19], Auto-Exposure [FZG*21], and DC-Shadow [JST21] produce obvious artifacts for shadow images with complex shadows, as shown in Figure 8(b), (c) and (d). How-

ever, the proposed method can effectively deal with these complicated shadows and produce better results, as shown in Figure 8(e).

4.6. Ablation study

To verify the benefits of each component in the proposed framework, we conduct ablation studies on the SRD dataset.

The proposed method is based on the theory that the geometry structure is an important factor in shadow generation. So we first conduct experiments to verify the specific impact of depth information on shadow removal through quantitative analysis. We use a black image with all pixels set to 0 as the prediction depth image, which indicates that the predicted depth map is not used in the experiment (*w/oD*). Moreover, we conduct experiments to verify the effectiveness of the boundary optimization strategy (*w/oB* indicates that we do not use the boundary refinement strategy). The effectiveness of each module is verified by comparing the values of RMSE, SSIM, and PSNR, as shown in Table 3.

From Table 3, we can see that the depth information plays an important role in the framework through the results in the first row. Especially for shadow areas, adding depth information improves the performance of RMSE from 5.37 to 4.59. Then, by comparing the data in the second row and the last row, we find that the proposed shadow boundary optimization strategy can improve the shadow area by reducing the RMSE from 5.08 to 4.59. Finally, by comparing the data in the third row and the last row, we verify the

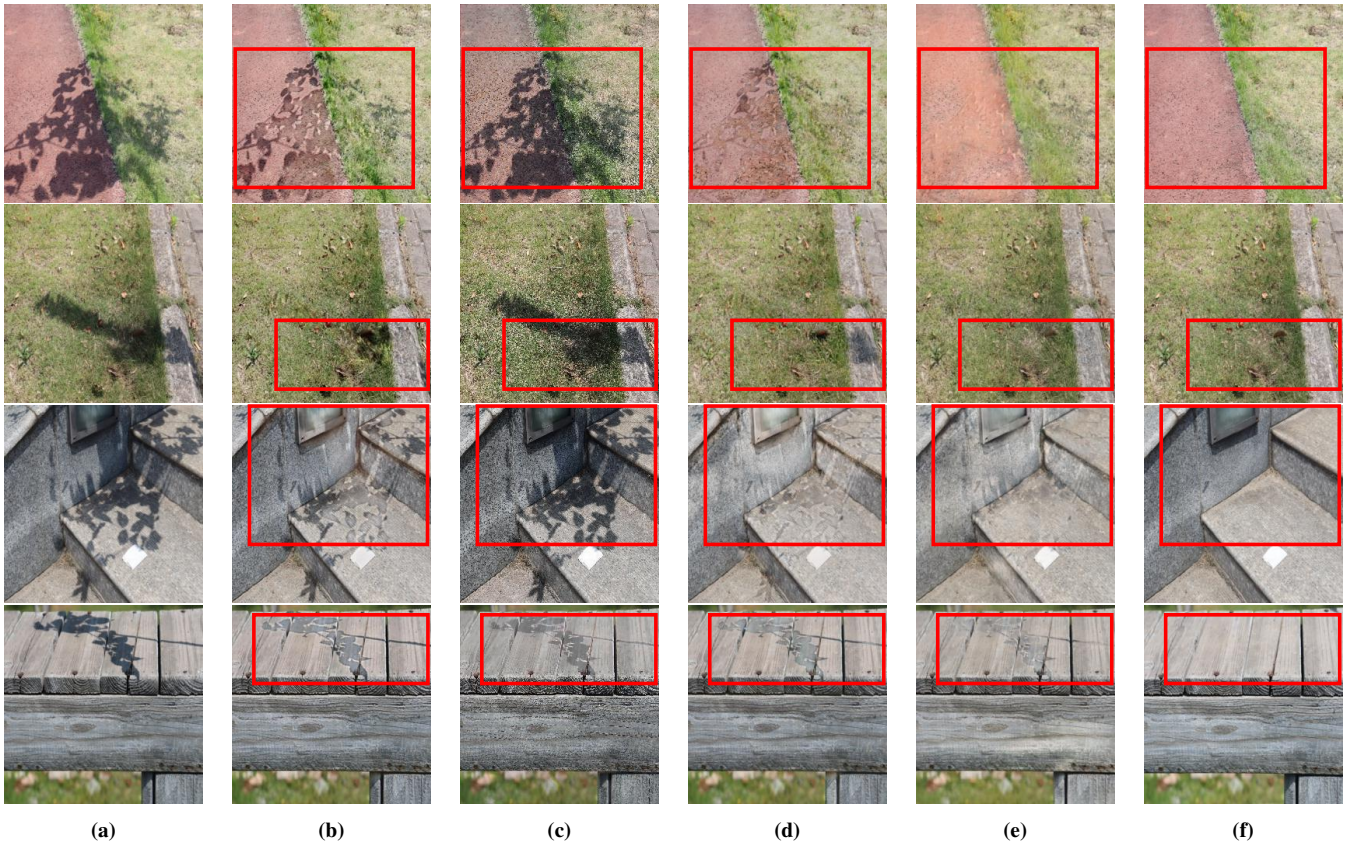


Figure 8: Shadow removal results of the real scenes captured by ourselves. (a) Input. (b) Mask-ShadowGAN [HJFH19]. (c) Auto-Exposure [FZG*21]. (d) DC-Shadow [JST21]. (e) Ours. (f) Ground Truth.

Table 3: Results of ablation studies on the SRD dataset.

Model	Shadow Region			Non-Shadow Region			All		
	SSIM \uparrow	PSNR \uparrow	RMSE \downarrow	SSIM \uparrow	PSNR \uparrow	RMSE \downarrow	SSIM \uparrow	PSNR \uparrow	RMSE \downarrow
(w/oD)	0.970	41.56	5.37	0.944	36.77	5.07	0.914	34.34	5.32
(w/oB)	0.981	41.82	5.08	0.950	37.02	3.69	0.921	34.62	4.02
Ours	0.982	42.48	4.59	0.960	38.69	3.38	0.934	36.19	3.69

impact on the results when the depth map estimation is inaccurate. When we add the predicted depth map as a guide, the performance of the proposed method is significantly improved. However, when we add gaussian noise to the predicted depth map to simulate the predicted depth errors, which demonstrates the effect of predicted depth error on performance.

Figure 9 shows the visual comparison results of the boundary optimization strategy. From Figure 9 we can find the effectiveness of the boundary refinement module. If we do not adopt the boundary optimization strategy, the results of shadow removal will have indistinct artifacts at the shadow boundaries, as shown in Figure 9(b). When we add the boundary optimization strategy, the artifacts at the shadow boundary will be significantly alleviated and close to Ground Truth, as shown in Figure 9(c).

4.7. Limitations

Our method uses depth information to guide the shadow removal, which can bring great benefits to shadow removal from the experimental results. However, the depth information is directly estimated from the shadow image by a depth prediction network. Therefore, if the training dataset is insufficient and the depth estimation error is significant, the predicted depth information with large errors may hurt the proposed shadow removal method.

5. Conclusion

In this work, we propose a novel depth-aware shadow removal framework without using depth images, which is composed of depth prediction, shadow removal, and boundary refinement. First, we design a depth prediction network to predict the depth of the

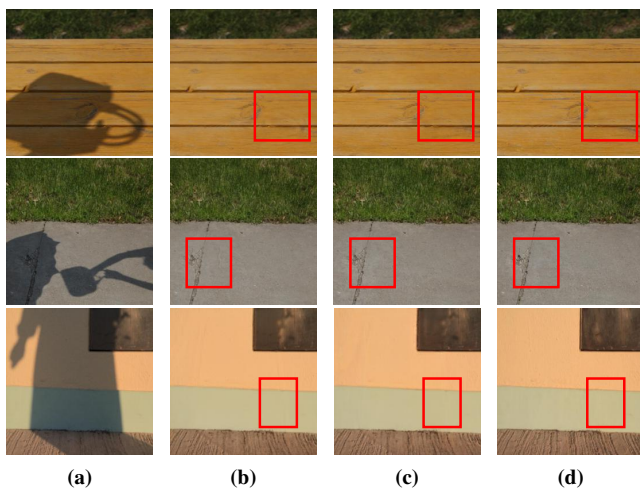


Figure 9: Visual comparison of ablation experiments for boundary optimization models. (a) Shadow images. (b) Without boundary refinement module. (c) With boundary refinement module. (d) Ground Truth.

shadow image. Second, we propose a shadow removal network based on the GAN architecture. For the generator, we design a dual encoder-decoder structure, one backbone for depth image feature extraction and the other for shadow image feature extraction. Then we introduce the multi-modality feature fusion and multi-scale feature fusion strategies to aggregate the complementary information between modules, which can effectively improve the performance of the proposed shadow removal framework. Finally, we propose a boundary refinement network using depth cues to refine the boundary within the penumbra region, which can further improve the results of shadow removal. Extensive experiments demonstrate that our method achieves superior shadow removal performance against state-of-the-art methods on public datasets and real-world scenes.

acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No.61876002, No.62076005, No.U20A20398), CCF-Tencent Open Fund (No. CCF-Tencent RAGR20210117), Anhui Natural Science Foundation Anhui Energy Internet Joint Fund (No.2008085UD07), National Natural Science Foundation of China and Anhui Provincial Key Research and Development Project (No.202104a07020029), The University Synergy Innovation Program of Anhui Province, China (No.GXXT-2021-030, No.GXXT-2021-002, No.GXXT-2021-065).

References

- [CCXH20] CHEN Z., CONG R., XU Q., HUANG Q.: DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection. *IEEE Transactions on Image Processing* (2020). 4
- [CGPP03] CUCCHIARA R., GRANA C., PICCARDI M., PRATI A.: Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25, 10 (2003), 1337–1342. 1

- [CLZX21] CHEN Z., LONG C., ZHANG L., XIAO C.: CANet: A context-aware network for shadow removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021). 2
- [CPS20] CUN X., PUN C.-M., SHI C.: Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2020). 6, 7
- [DC19] DIJK T. V., CROON G. D.: How do neural networks see depth in single images? In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 2183–2191. 3
- [EK17] ENGEL J., KOLTUN V., CREMERS D.: Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 3 (2017), 611–625. 1
- [FDL09] FINLAYSON G. D., DREW M. S., LU C.: Entropy minimization for shadow removal. *International Journal of Computer Vision* 85, 1 (2009), 35–57. 2
- [FHL05] FINLAYSON G. D., HORDLEY S. D., LU C., DREW M. S.: On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 1 (2005), 59–68. 2
- [FWZ*20] FAN X., WU W., ZHANG L., YAN Q., FU G., CHEN Z., LONG C., XIAO C.: Shading-aware shadow detection and removal from a single image. *The Visual Computer* 36, 10 (2020), 2175–2188. 2
- [FZG*21] FU L., ZHOU C., GUO Q., JUEFEI-XU F., YU H., FENG W., LIU Y., WANG S.: Auto-exposure fusion for single-image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021). 1, 2, 6, 7, 8
- [GC14] GONG H., COSKER D.: Interactive shadow removal and ground truth for variable scene categories. In *British Machine Vision Conference* (2014). 6, 7
- [GC17] GONG H., COSKER D.: User-assisted image shadow removal. *Image and Vision Computing* 62 (2017), 19–27. 2
- [GDH12] GUO R., DAI Q., HOIEM D.: Paired regions for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 12 (2012), 2956–2967. 7
- [HFY20] HAN Z., FU Z., YANG J.: Learning the redundancy-free features for generalized zero-shot object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020). 1
- [HFZ*19] HU X., FU C.-W., ZHU L., QIN J., HENG P.-A.: Direction-aware spatial context features for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 11 (2019), 2795–2808. 2, 6, 7
- [HJFH19] HU X., JIANG Y., FU C.-W., HENG P.-A.: Mask-ShadowGAN: Learning to remove shadows from unpaired data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019). 2, 6, 7, 8
- [HZW*21] HU X., ZHU L., WANG T., FU C.-W., HENG P.-A.: Single-image real-time rain removal based on depth-guided non-local features. *IEEE Transactions on Image Processing* 30 (2021), 1759–1770. 2
- [JST21] JIN Y., SHARMA A., TAN R. T.: Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 5027–5036. 1, 2, 5, 6, 7, 8
- [LCC20] LIN Y.-H., CHEN W.-C., CHUANG Y.-Y.: BEDSR-Net: A deep shadow removal network from a single document image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020). 2
- [LS19] LE H., SAMARAS D.: Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019). 7
- [LS20] LE H., SAMARAS D.: From shadow segmentation to shadow removal. In *European Conference on Computer Vision* (2020), pp. 264–281. 2, 7

- [LYM*21] LIU Z., YIN H., MI Y., PU M., WANG S.: Shadow removal by a lightness-guided network with training on unpaired data. *IEEE Transactions on Image Processing* 30 (2021), 1853–1865. 2, 7
- [LYW*21] LIU Z., YIN H., WU X., WU Z., MI Y., WANG S.: From shadow generation to shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021). 1, 2, 3, 7
- [MAT17] MUR-ARTAL R., TARDÓS J. D.: ORB-SLAM2: An open-source slam system for monocular, stereo, and RGB-D cameras. *IEEE transactions on robotics* 33, 5 (2017), 1255–1262. 1
- [MGK21] MURALI S., GOVINDAN V., KALADY S.: Quaternion-based image shadow removal. *The Visual Computer* (2021), 1–12. 2
- [NB04] NADIMI S., BHANU B.: Physical models for moving shadow and object detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, 8 (2004), 1079–1087. 1
- [QCHX19] QU Y., CHEN Y., HUANG J., XIE Y.: Enhanced pix2pix dehazing network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019). 5
- [QTH*17] QU L., TIAN J., HE S., TANG Y., LAU R. W.: Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017). 2, 5, 6
- [QYL*21] QIAN Y., YAN S., LUKEŽIČ A., KRISTAN M., KÄMÄRÄINEN J.-K., MATAS J.: DAL: A deep depth-aware long-term tracker. In *International Conference on Pattern Recognition* (2021). 2
- [RBK21] RANFTL R., BOCHKOVSKIY A., KOLTUN V.: Vision transformers for dense prediction. 12179–12188. 3
- [SHKF12] SILBERMAN N., HOIEM D., KOHLI P., FERGUS R.: Indoor segmentation and support inference from RGB-D images. In *European conference on computer vision* (2012), pp. 746–760. 3
- [SL08] SHOR Y., LISCHINSKI D.: The shadow meets the mask: Pyramid-based shadow removal. In *Computer Graphics Forum* (2008), vol. 27, pp. 577–586. 2
- [VHS17] VICENTE T. F. Y., HOAI M., SAMARAS D.: Leave-one-out kernel optimization for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 3 (2017), 682–695. 2
- [VRVGT21] VASLUIANU F.-A., ROMERO A., VAN GOOL L., TIMOFTE R.: Shadow removal with paired and unpaired learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021). 2
- [WLY18] WANG J., LI X., YANG J.: Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018). 2, 3, 5, 7
- [WSL*20] WANG L.-W., SIU W.-C., LIU Z.-S., LI C.-T., LUN D. P.: Deep relighting networks for image light source manipulation. In *European Conference on Computer Vision* (2020). 1
- [WTBS07] WU T.-P., TANG C.-K., BROWN M. S., SHUM H.-Y.: Natural shadow matting. *ACM Transactions on Graphics (TOG)* 26, 2 (2007), 8–es. 2
- [XRG*17] XIE Q., REMIL O., GUO Y., WANG M., WEI M., WANG J.: Object detection and tracking under occlusion for object-level RGB-D video segmentation. *IEEE Transactions on Multimedia* 20, 3 (2017), 580–592. 1
- [XTT14] XIAO Y., TSOUGENIS E., TANG C.-K.: Shadow removal from single RGB-D images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014). 1, 2
- [XZW*20] XIAN K., ZHANG J., WANG O., MAI L., LIN Z., CAO Z.: Structure-guided ranking loss for single image depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 611–620. 3
- [YME*20] YU Y., MEKA A., ELGHARIB M., SEIDEL H.-P., THEOBALT C., SMITH W. A.: Self-supervised outdoor scene relighting. In *European Conference on Computer Vision* (2020). 1
- [ZLZX20] ZHANG L., LONG C., ZHANG X., XIAO C.: RIS-GAN: Explore residual and illumination with generative adversarial networks for shadow removal. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2020). 2
- [ZXF*22] ZHU Y., XIAO Z., FANG Y., FU X., XIONG Z., ZHA Z.-J.: Efficient model-driven network for shadow removal. 2, 6, 7
- [ZZJ*21] ZHANG Y.-F., ZHENG J., JIA W., HUANG W., LI L., LIU N., LI F., HE X.: Deep RGB-D saliency detection without depth. *IEEE Transactions on Multimedia* (2021). 2
- [ZZX15a] ZHANG L., ZHANG Q., XIAO C.: Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing* 24, 11 (2015), 4623–4636. 2
- [ZZX15b] ZHANG L., ZHANG Q., XIAO C.: Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing* 24, 11 (2015), 4623–4636. 2