MGRLN-Net: Mask-Guided Residual Learning Network for Joint Single-Image Shadow Detection and Removal

Leiping Jie[1,2][0000−0002−1681−7926] and Hui Zhang[2][0000−0001−9387−8477](✉)

[1] Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[2] Department of Computer Science and Technology, BNU-HKBU United International College, Zhuhai, China
amyzhang@uic.edu.cn

Abstract. Although significant progress has been made in single-image shadow detection or single-image shadow removal, only few works consider these two problems together. However, the two problems are complementary and can benefit from each other. In this work, we propose a Mask-Guided Residual Learning Network (MGRLN-Net) that jointly estimates shadow mask and shadow-free image. In particular, MGRLN-Net first generates a shadow mask, then utilizes a feature reassembling module to align the features from the shadow detection module to the shadow removal module. Finally, we leverage the learned shadow mask as guidance to generate a shadow-free image. We formulate shadow removal as a masked residual learning problem of the original shadow image. In this way, the learned shadow mask is used as guidance to produce better transitions in penumbra regions. Extensive experiments on ISTD, ISTD+, and SRD benchmark datasets demonstrate that our method outperforms current state-of-the-art approaches on both shadow detection and shadow removal tasks. Our code is available at `https://github.com/LeipingJie/MGRLN-Net`.

Keywords: Shadow detection and removal · Multi-task learning · Masked residual learning.

1 Introduction

Shadows that help us better understand real-world scenes are cast by objects that block the propagation of light rays and are ubiquitous in our daily lives. However, they cause trouble to many tasks, *e.g.*, object detection, image segmentation, or scene analysis. Shadows can be cast into arbitrary shapes with different intensities at any position, making both shadow detection and removal challenging.

Due to their challenge and importance, shadow detection and removal are active research topics. Traditional methods [30,6,5,18] utilizing physical models,

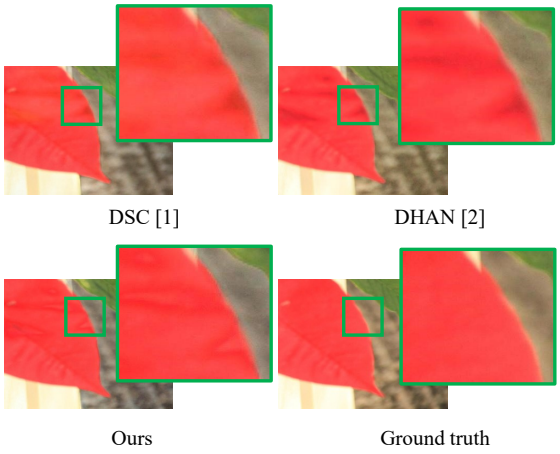

Fig. 1: We compare our model with two state-of-the-art methods DSC [11] and DHAN [3]. As can be seen, our model produces better details.

handcrafted features, or prior knowledge, are not robust and lead to unsatisfactory performance. Leveraging large-scale annotated datasets and computational power, deep learning-based shadow detection and shadow removal approaches have shown their superiority. Modern shadow detection methods [35,21,37,31,14] formulate the shadow detection problem as a binary classification problem. They typically use different strategies to extract the global and local contexts from a single input image, including attention mechanisms [15,4], bidirectional fusion [37], and teacher-student learning [1]. Similarly, shadow removal approaches [28,12] primarily leverage Generative Adversarial Networks (GAN), where the generator attempts to produce faked shadow-free images while the discriminator tries to distinguish between the real shadow-free images and the generated fake shadow-free images. Despite superior performance, most approaches tackle these two problems individually. Intuitively, shadow detection and shadow removal are mutually beneficial. On the one hand, the shadow detection results provide strong guidance for shadow removal algorithms to adjust more on shadow pixels and less on non-shadow pixels. On the other hand, the shadow removal process expects more variation on shadowed pixels and less variation on non-shadowed pixels, which is also associated with identifying whether a pixel is a shadow pixel or not.

In this paper, we propose a unified network for joint shadow detection and removal. In practice, shadow detection and removal are formulated as classification and regression problems. For the shadow detection problem, the model only needs to predict whether a pixel is a shadow pixel or not. However, the model needs to answer how to transfer shadowed pixels to shadow-free pixels to solve the shadow removal problem, which is more challenging. Based on this, we design a compact and efficient shadow detection sub-network and a more

complicated shadow removal sub-network. Specifically, we build our model in an encoder-decoder way, where the encoder extracts multi-level features while the decoder is responsible for fusing the features to generate the desired shadow masks and shadow-free images. We want to emphasize that although several papers [11,4,13] claimed for performing joint shadow detection and removal, they differ from ours. [11,13] introduced frameworks that could be trained for shadow detection or shadow removal. In other words, their frameworks were trained separately for shadow detection and removal. However, our network can be trained with the two tasks concurrently. [4] differs from ours in five aspects: (1) Their architecture is based on GAN. (2) They do not explicitly predict shadow mask images. (3) They use the recurrent unit to generate shadow attention, while we use it to remove shadows. (4) Our recurrent unit is shared while theirs is not. (5) Our model performs better on shadow detection and removal.

One challenging problem for shadow removal is to generate a natural transition effect in the penumbra regions (between the umbra and the shadow-free regions). Nevertheless, identifying and annotating penumbra regions is exceptionally time-consuming, expensive, or even impossible. To generate annotation for penumbra regions at low cost, previous methods [7] utilize morphological algorithms. Specifically, the penumbra region is defined as the area of the dilation mask minus the erosion mask. However, the kernels used for dilation and erosion are chosen empirically and thus are sometimes inappropriate. Thanks to our joint training pipeline, the shadow masks predicted by our network naturally contain penumbra regions. In other words, the shadow boundaries in the shadow mask do not transition hardly but softly. Consequently, we design a shadow mask-guided residual learning module to remove shadows. Specifically, the shadow-guided residual learning module consists of feature reassembling, feature refining, and prediction modules.

To verify the effectiveness of our proposed method, we conduct extensive experiments on three commonly used benchmark datasets: ISTD, ISTD+, and SRD. Comparisons are made with both state-of-the-art shadow detection and removal methods. Experimental results show that our network outperforms the state-of-the-art shadow detection and removal methods.

In summary, our main contributions are three-fold:

- We propose a novel network for joint shadow detection and shadow removal.
- We design an efficient shadow removal module that reassembles and refines the context features with the mask guidance to produce better transition effects in penumbra regions.
- We show that the proposed network outperforms the state-of-the-art shadow detection and removal methods on three widely used benchmark datasets ISTD [28], SRD [22] and ISTD+ [19].

2 Related Work

In this section, we review the shadow detection and the shadow removal approaches, respectively.

2.1 Shadow Detection

Before the era of deep learning, most shadow detection approaches relying on the physical properties [6,5] assumed color, illumination or statistical-based handcraft features to be consistent [36,18]. Zhu *et al.* [36] combined mixed features, *e.g.*, the intensity difference, the gradient, and the texture similarities, to train a boosted decision tree classifiers. For better performance and robustness, Guo *et al.* [10] considered areas rather than individual pixels or edges to construct a graph of segments by classifying the pairwise segmented area, followed by the graph-cut algorithm. Later, Vicente *et al.* [27] distinguished the shadow area from the non-shadow region by training a kernel Least-Squares Support Vector Machine (LSSVM). Despite the improved performance, traditional approaches heavily rely on consistent color or illumination assumptions, which may not be suitable for real-world scenes. Therefore, the overall performance is not high and satisfactory. Like many other computer vision tasks, shadow detection is now dominated by deep learning approaches. Early solutions utilized the deep convolutional neural network (CNN) as the feature extractor to replace handcraft designs. Khan *et al.* [17] proposed the first CNN-based approach for shadow detection. Unlike traditional methods, they used a 7-layer CNN to learn features along the object boundaries at a super-pixel level and then generated smooth shadow contours with a conditional random field model. Shen *et al.* [24] exploited the local structure of the shadow edge using the structured CNN and improved the local consistency of the estimated shadow map with the structured labels. Later, due to the newly developed neural network architectures, such as U-Net [23], GAN [8], research on shadow detection tended to train neural networks in an end-to-end manner, focusing more on using both global and local features at the same time. Hu *et al.* [31] demonstrated that the direction-aware context features could be learned by spatial recurrent neural network (RNN). Zhu *et al.* [37] presented a bidirectional feature pyramid network to explore and combine global and local context. Recently, Jie *et al.* [14,15] proposed a transformer-based network to capture attention along multi-level features. Unlike the common CNN architecture as the feature extractor, Nguyen *et al.* [21] added an additional parameter of sensitivity to the generator to optimize the based conditional GAN framework. More recently, Zheng *et al.* [35] proposed to explicitly learn and integrate the semantics of visual distraction areas with their differentiable Distraction-aware Shadow (DS) module. To alleviate the burden of annotation and boost performance, Chen *et al.* [1] introduced to explore the learning of multiple information of shadows using a multi-task mean teacher model with unlabeled data in a semi-supervised manner.

2.2 Shadow Removal

Similar to the shadow detection problem, shadow removal methods can also be divided into traditional and learning-based methods. Traditional approaches exploited physical properties, *e.g.*, image illumination [32,33] and image gradient [9]. Recently, shadow removal approaches using deep learning have become popular. Qu *et al.* [22] introduced a multi-context architecture to integrate

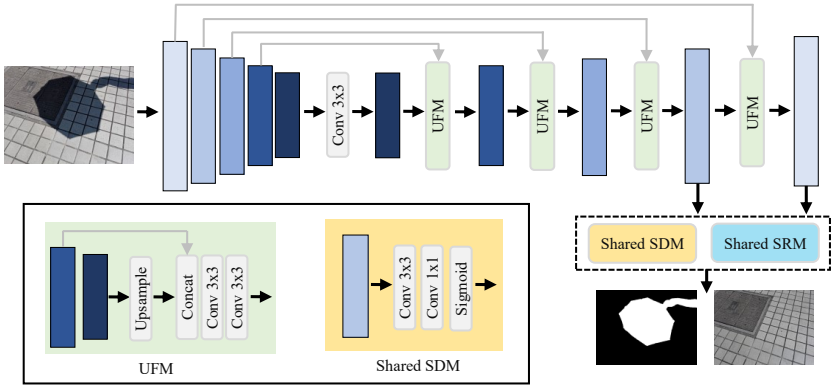

Fig. 2: Overview of our proposed network. Our network takes a single image as input and outputs the corresponding shadow mask and shadow-free image. The network is an encoder-decoder structure and is composed of a feature extractor (see Section 3.1), a shared shadow detection module (see Section 3.2) and a shared shadow removal module (see Fig. 3 and Fig. 4). The input image is fed into the feature extractor to obtain multi-level feature maps, which are fed into both the shadow detection and shadow removal module. The predicted shadow mask is also fed into the shadow removal module as weight guidance for the residual learning of the shadow-free images.

high-level semantic context, mid-level appearance information, and local image details, which learns a mapping function between the shadow image and its shadow matte. Hu *et al.* [11] leverages the spatial context in different directions and attention mechanism for both shadow detection and removal. Chen *et al.* [2] proposed a two-stage context network to transfer contextual information from non-shadow patches to shadow patches. By formulating the shadow removal problem as an exposure fusion problem, Fu *et al.* [7] addressed shadow removal by fusing estimated over-exposure images and achieved state-of-the-art performance. Moreover, methods based on Generative Adversarial Network (GAN) show their potential. ST-CGAN [28] utilized two conditional GAN for both shadow detection and removal. MaskShadowGAN [12] proposed a cycle-GAN-based framework to identify the shadow-free to shadow image translation with learned guidance from shadow masks. Despite the boosting performance, only few works consider both shadow detection and shadow removal. We believe that the two complementary tasks can benefit from each other and should be considered together.

3 METHODOLOGY

In this section, we first introduce the overall architecture of our proposed method in Section 3.1. Next, we illustrate our shadow detection module in Section 3.2 and

shadow removal module in Section 3.3, respectively. Finally, the loss functions will be presented in Section 3.4.

3.1 Network Architecture

As shown in Fig. 2, our method takes a single shadow image as input and outputs the corresponding shadow mask and shadow-free image. Specifically, we first utilize a pretrained EfficientNet [26] as our feature extractor, which obtains L different levels of encoder features $\{F_i\}_{i=1}^{L}$ ($L = 5$ here). These features are then fed into the following modules, including an Upsampling and Fusion Module (UFM), a Shadow Detection Module (SDM), and a Shadow Removal Module (SRM), to generate different levels of decoder features $\{D_i\}_{i=1}^{L}$ ($L = 5$ here), where F_i and D_i are of the same resolution. In UFM, it first upsamples the current decoder feature map $D_i (i \neq 0)$ to D_i' using a differentiable interpolation operator, and then concatenates D_i' with the corresponding encoder feature map F_{i-1}, followed by two consecutive 3×3 convolutional layers. We denote the output feature map of UFM as $\{U_i\}_{i=1}^{L}$ ($L = 4$ here).

3.2 Shadow Detection Module

Considering that our shadow detection module is used to help with shadow removal, we designed a compact and efficient subnetwork for shadow detection. Given any output feature map U_i from UFM, we stack a 3×3 and another 1×1 convolutional layer with the sigmoid function to generate the predicted shadow mask. In our model, the output feature map U_i from UFM is also fed into the shadow removal module, which implies that U_i contains the discriminative features for both shadow detection and removal when training is processed in a joint manner. Unlike previous methods that generate shadow masks from each decoder layer, we only generate two shadow masks from the last two decoder layers. Despite its simple structure, our shadow detection module predicts satisfactory shadow masks (more details are presented in Section 4.3)

3.3 Shadow Removal Module

As shown in Fig. 3, our proposed shadow removal module consists of three submodules: Feature Reassembling (FR), Recurrent Refinement (RR), and Residual Learning (RL).

Feature Reassembling. The feature reassembling submodule aims at reassembling the input feature U_i. Since our shadow detection module and shadow removal module share U_i, it is difficult to force U_i to be discriminative for both shadow detection and shadow removal. We argue that discriminative features for shadow detection and removal should be different but complementary, which means they can be transferred from one kind to the other. Based on this, we adopt an improved lightweight U-Net to accomplish this transformation. The

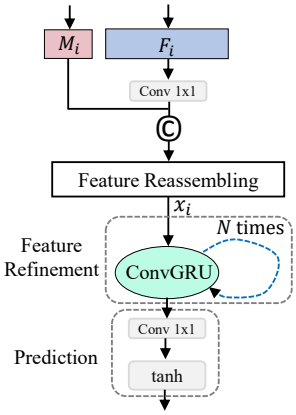

Fig. 3: Illustration of our proposed shadow removal module.

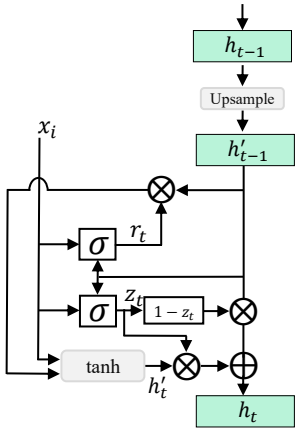

Fig. 4: Illustration of ConvGRU module.

UNet downsamples and upsamples three times, respectively, and keeps the input and output feature maps at the same resolution.

Recurrent Refinement. Although the reassembled features are adapted for shadow removal, they still contain noises. Here we propose to remove them in a recurrent way. Specifically, the recurrent refinement submodule utilizes ConvGRU [25]. Given the input feature map x_i and the previous hidden state h_{t-1}, we first apply two 3×3 convolution layers on x_i and h_{t-1}, respectively, followed by a sigmoid function to get the update gate z_t. The same operation is performed to get the reset gate r_t. Then, r_t is used to generate candidate hidden state h'_t. Finally, the update gate z_t is used to adaptively select information from the previous hidden state h_{t-1} and the candidate hidden state h'_t to output h_t. The whole procedure can be formulated as follows:

$$\begin{aligned} z_t &= \sigma(Conv_z^x(x_i) + Conv_z^h(h_{t-1})), \\ r_t &= \sigma(Conv_r^x(x_i) + Conv_r^h(h_{t-1})), \\ h'_t &= tanh(Conv_h^x(x_i) + Conv_h^r(r_t \circ h_{t-1})), \\ h_t &= (1 - z_t) \circ h_{t-1} + z_t \circ h'_t, \end{aligned} \tag{1}$$

where ©, Conv, σ, $\circ$ are concatenation operation, convolutional layer, sigmoid function and element-wise multiplication, respectively. It is worth mentioning that the recurrent refinement can be run N times. Empirically, we set $N = 2$ for speed-performance tradeoff.

Residual Learning. To get the final shadow-free image prediction, we use residual learning, which means we regress the residual image R_i based on the input RGB image instead of predicting the shadow-free image directly. This is quite effective in our experiment (see Section 4.5). More importantly, we impose the predicted shadow mask M_i on the residual output, which enables our network to generate desired transition effect between the shadow-free and penumbra

regions. We express this procedure as follows:

$$Free = M_i \circ R_i + RGB, \tag{2}$$

where $Free$, RGB, M_i, R_i represent the predicted shadow-free image, the input RGB shadow image, the predicted shadow mask, and the predicted residual image, respectively.

3.4 Objective Functions

For both predicted shadow mask and shadow-free images, we use $L1$ loss as follows:

$$\begin{aligned} L_{rgb} &= \sum_{i=1}^{2} \|y_i - \hat{y}_i\|_1, \\ L_{mask} &= \sum_{i=1}^{2} \|m_i - \hat{m}_i\|_1, \end{aligned} \tag{3}$$

where m_i, $\hat{m}_i$, y_i, $\hat{y}_i$ are the predicted shadow mask, the ground truth shadow mask, the predicted shadow-free image, and the ground truth shadow-free image, respectively.

Furthermore, we also compute the feature loss using the pretrained *VGG-19* network Ψ as follows:

$$L_{feature} = \sum_{l=1}^{\Omega} \|\Psi_l(y) - \Psi_l(\hat{y})\|_1, \tag{4}$$

where Ψ_l indicates the layer l, and Ω represents the $3th, 8th, 15th, 22th$ layers in the *VGG-19* network.

Overall, our loss function is:

$$L = L_{mask} + \lambda_1 L_{rgb} + \lambda_2 L_{feature}, \tag{5}$$

where λ_1 and λ_2 are empirically set to 2.0 to balance between different losses.

4 Experimental Results

4.1 Datasets and Evaluation Metrics

Datasets. We train and evaluate our proposed method on three widely used benchmark datasets: ISTD [28], ISTD+ [19] and SRD [22]. ISTD consists of $1,300$ and 540 triplets of shadow, shadow mask, and shadow-free images for training and testing. ISTD+ is constructed based on the ISTD, where only the shadow-free images are adjusted for color consistency between shadow and shadow-free images, and thus has the same training and testing splits as ISTD. In

contrast, SRD has 2,680 and 408 shadow and shadow-free image pairs for training and testing, with no shadow masks provided. Since shadow masks are necessary for our pipeline, we follow MaskShadow-GAN [12] to generate shadow masks by using Otsu's algorithm with the difference between shadow and shadow-free images. The image resolution in ISTD and ISTD+ is 640×480, while SRD is 840×640.

Evaluation Metric. We employ the Balance Error Rate (BER) and Root Mean Square Error ($RMSE$) to quantitatively evaluate shadow detection and shadow removal performance. BER considers the performance of both shadow prediction and non-shadow prediction and can be formulated as follows:

$$BER = \left(1 - \frac{1}{2}\left(\frac{TP}{N_p} + \frac{TN}{N_n}\right)\right) \times 100 \ , \tag{6}$$

where TP, TN, N_p, and N_n are the number of true positive pixels, true negative pixels, shadow pixels, and non-shadow pixels, respectively. $RMSE$ is calculated in the LAB color space between the predicted shadow-free images and the ground truth shadow-free image. It is worth noting that the default evaluates code used by all methods (including ours) actually computes the mean absolute error (MAE), as mentioned in [16,19]. For both BER and $RMSE$, the smaller the value, the better the performance.

4.2 Implementation details

Our proposed method is implemented in PyTorch and all the experiments are conducted on a NVIDIA single RTX 2080Ti GPU.

Training Settings. When training, we crop and resize the input images to 448×448 with batch size 8. The maximum learning rate max_{lr} is set to 0.000375 and decayed with 1-cycle policy. Specifically, the initial and the minimum learning rate are set to $max_{lr}/30$ and $max_{lr}/150$, while the percentage of the cycle spent increasing the learning rate is 0.1. We empirically train our network for 200 epochs using half-precision floating and AdamW optimizer, where the first momentum value, the second momentum value, and the weight decay are 0.9, 0.999, and 5^{-4}, respectively.

Testing Setting. When testing, we do not apply any data augmentations and post-processing operations, *e.g.*, conditional random filed (CRF) for shadow masks.

4.3 Comparison with State-of-the-Art Shadow Detection Methods

Since no ground truth shadow mask is provided for SRD, we only evaluate our shadow detection results on ISTD. As shown in Table 1, our model achieves the best performance against ST-CGAN [29], DSDNet [35], BDRAR [37] and DSC [11]. It is worth mentioning that DSDNet, BDRAR, and DSC are elaborately designed for the shadow detection task only. Nevertheless, we can still outperform DSDNet, DSC and BDRAR by 33.18%, 46.1% and 57.6%. More

importantly, compared with the existing jointly training frameworks ST-CGAN and ARGAN for shadow detection and removal, our performance surpasses theirs significantly by 83.14% and 27.86%.

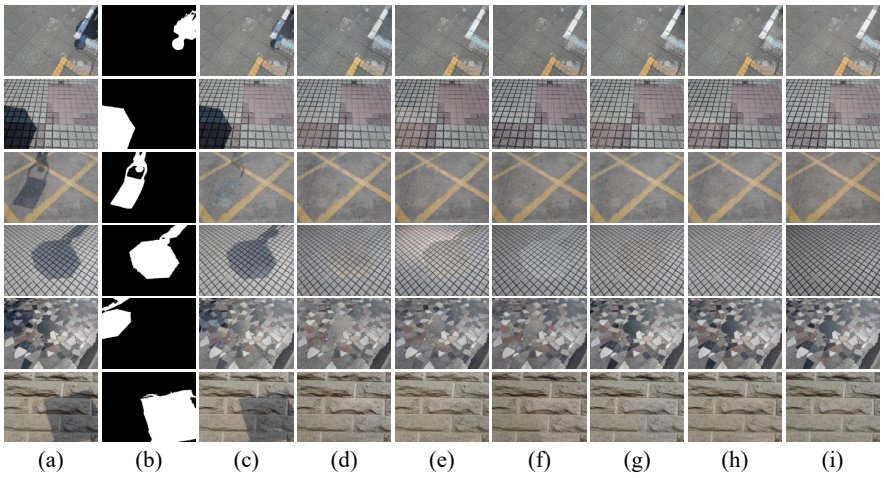

Fig. 5: Qualitative comparison on ISTD [28] dataset. From left to right: (a) input shadow image; (b) shadow mask; (c) Guo *et al.* [10]; (d) ST-CGAN [29]; (e) DSC [11]; (f) DHAN [3]; (g) Auto-Exposure [7]; (h) ours; (i) ground truth shadow-free image. Best viewed on screen.

4.4 Comparison with State-of-the-art Shadow Removal Methods

We compare the performance of our network with different state-of-the-art methods since some only provide their performance on one or two benchmark datasets.
Comparison on ISTD Dataset. On the ISTD dataset, we compare with 8 other methods: Guo *et al.* [10], Zhang *et al.* [34], MaskShadow-GAN [12], ST-CGAN [29], DSC [11], DHAN [3], CANet [2], Auto-Exposure [7]. As can be seen from Table 2, our method achieves the best performance. In particular, our method outperforms Guo *et al.* [10], Zhang *et al.* [34], MaskShadow-GAN [12], ST-CGAN [29], DSC [11], DHAN [3], CANet [2], Auto-Exposure [7] by 45.91%, 41.03%, 32.12%, 32.66%, 24.59%, 21.04%, 18.21%, 15.03%, respectively. We also visualize our predicted shadow-free images in Fig. 5. Qualitatively our method generates satisfactory predictions on the ISTD benchmark dataset.
Comparison on the ISTD+ Dataset. On the ISTD+ dataset, we compare with 8 other methods: Guo *et al.* [10], Zhang *et al.* [34], ST-CGAN [29], DeshadowNet [22], MaskShadow-GAN [12], Param+M+D-Net [20], SP+M-Net [19], Auto-Exposure [7]. As presented in Table 3, our method achieves the best performance. Specifically, our model surpasses Guo *et al.* [10], ST-CGAN [29], Deshad-

Table 1: Quantitative comparison of shadow detection performance on ISTD [28]. The best and the second best results are highlighted in bold and underlined, respectively.

Method	BER	Shadow	Non-Shadow
ST-CGAN [29]	8.60	7.69	9.23
DSDNet [35]	<u>2.17</u>	<u>1.36</u>	<u>2.98</u>
BDRAR [37]	2.69	**0.50**	4.87
DSC [11]	3.42	3.85	3.00
ARGAN [4]	2.01	-	-
Ours	**1.45**	1.65	**1.26**

Table 2: Quantitative comparison of shadow removal on ISTD [28]. The best and the second best results are highlighted in bold and underlined, respectively.

Method / RMSE	Shadow	Non-Shadow	All
Input Image	32.12	7.19	10.97
Guo *et al.* [10]	18.95	7.46	9.30
Zhang *et al.* [34]	14.98	7.29	8.53
MaskShadow-GAN [12]	12.67	6.68	7.41
ST-CGAN [29]	10.33	6.93	7.47
DSC [11]	9.76	6.14	6.67
DHAN [3]	8.14	6.04	6.37
CANet [2]	8.86	6.07	6.15
Auto-Exposure [7]	<u>7.77</u>	<u>5.56</u>	<u>5.92</u>
Ours	**7.65**	**4.52**	**5.03**

Table 3: Quantitative comparison of shadow removal on ISTD+ [19]. The best and the second best results are highlighted in bold and underlined, respectively.

Method / RMSE	Shadow	Non-Shadow	All
Input Image	40.2	2.6	8.5
Guo *et al.* [10]	22.0	3.1	6.1
Zhang *et al.* [34]	13.3	-	-
ST-CGAN [29]	13.4	7.7	8.7
DeshadowNet [22]	15.9	6.0	7.6
MaskShadow-GAN [12]	12.4	4.0	5.3
Param+M+D-Net [20]	9.7	<u>3.0</u>	4.0
SP+M-Net [19]	7.9	3.1	<u>3.9</u>
Auto-Exposure [7]	**6.5**	3.8	4.2
Ours	<u>6.69</u>	**2.46**	**3.15**

Table 4: Quantitative comparison of shadow removal on SRD [22]. The best and the second best results are highlighted in bold and underlined, respectively.

Method / RMSE	Shadow	Non-Shadow	All
Input Image	40.28	4.76	14.11
Guo *et al.* [10]	29.89	6.47	12.60
DeshadowNet [22]	11.78	4.84	6.64
Auto-Exposure [7]	8.56	5.75	6.51
DSC [11]	10.89	4.99	6.23
CANet [2]	**7.82**	5.88	5.98
DHAN [3]	8.94	<u>4.80</u>	<u>5.67</u>
Ours	<u>8.03</u>	**3.27**	**4.93**

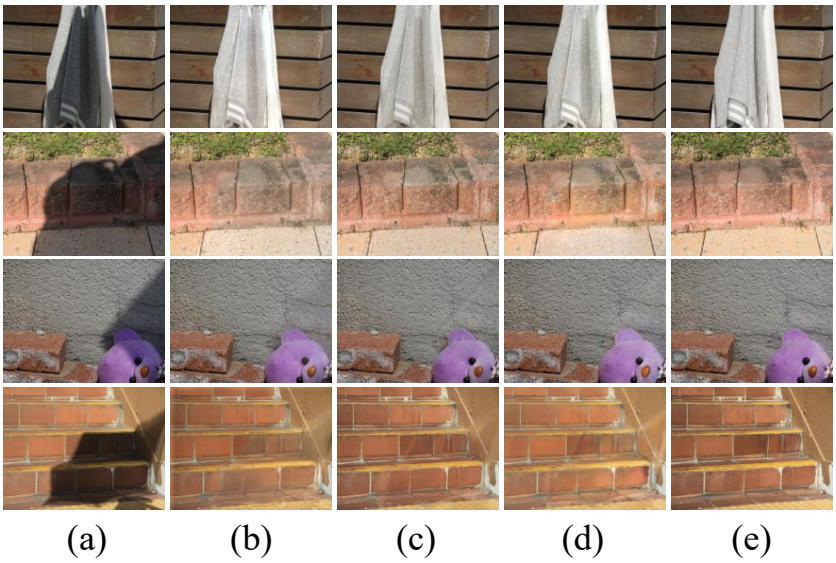

Fig. 6: Qualitative comparison on SRD [22] dataset. From left to right: (a) input shadow image; (b) DSC [11]; (c) DHAN [3]; (d) ours; (e) ground truth shadow-free image. Best viewed on screen.

owNet [22], MaskShadow-GAN [12], Param+M+D-Net [20], SP+M-Net [19], Auto-Exposure [7] by 48.36%, 63.8%, 58.55%, 40.57%, 21.25%, 19.23%, 25%, respectively.

Comparison on the SRD Dataset. On the SRD dataset, we compare with 6 other methods: Guo *et al.* [10], DeshadowNet [22], Auto-Exposure [7], DSC [11], CANet [2], DHAN [3]. As shown in Table 4, our method achieves the best performance. Quantitatively, our model outperforms Guo *et al.* [10], DeshadowNet [22], Auto-Exposure [7], DSC [11], CANet [2], DHAN [3] by 60.87%, 25.75%, 24.27%,

Table 5: Ablation studies of modules in our model.

baseline	RL	FR	RR	All
✓	×	×	×	7.95
✓	✓	×	×	6.43
✓	×	✓	×	5.93
✓	✓	✓	×	5.80
✓	✓	×	✓	5.32
✓	✓	✓	✓	**5.03**

Table 6: Ablation study on hyper-parameter N in the recurrent refinement module.

N	Shadow	Non-Shadow	All
1	8.17	4.88	5.42
2	**7.65**	**4.52**	**5.03**
3	7.95	4.62	5.16
4	8.11	4.56	5.14

20.87%, 17.56%, 13.05%, respectively. Meanwhile, we also produce qualitatively satisfactory shadow-free predictions. As illustrated in Fig. 6, our method can predict a better consistent appearance as the ground truth shadow-free image.

4.5 Ablation Studies

To evaluate the effectiveness of our proposed modules and the impact of different hyperparameter settings, we conduct an extensive ablation study in this section.

Fig. 7: Shadow removal results. From left to right are: prediction of the baseline model, prediction of the baseline with FR submodule, prediction of baseline with FR and RL submodules, prediction of the proposed model with all submodules, and ground truth shadow-free image.

Effectiveness of our network. To deeply analyze how different components affect performance, we first train a baseline model which only contains the feature extractor, the upsampling fusion block. Then we gradually add the Residual Learning (RL), the Feature Reassembling module (FR), and the Recurrent Refinement module (RR). As can be seen from Table 5, the RL module performs the best in terms of performance improvement. Every submodule in the shadow removal module is positive, and we achieve the best performance with all of them. It can be also seen from Fig. 7 that as FR, RR, and RL are gradually equipped with the baseline model, our predicted shadow-free images continue to improve.

Settings of Times N. Our recurrent refinement module contains a ConvGRU unit, which means we can run it recurrently without adding more training parameters. However, more running loops will lead to more time-consuming. We

choose four different values $N = 1, 2, 3, 4$. As shown in Table 6, $N = 2$ achieves the best performance, but $N = 3$ and $N = 4$ show nearly the same performance.

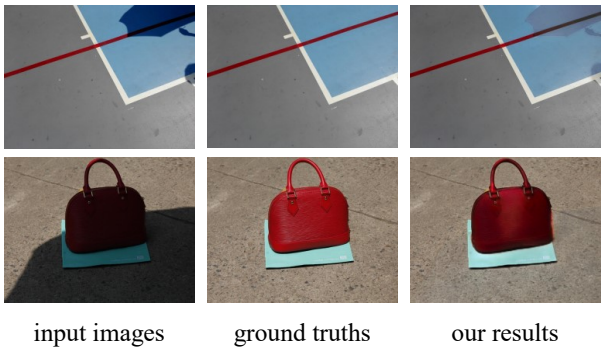

Fig. 8: Failure cases on shadow removal.

4.6 Failure Cases

Despite the superior performance, our method fails with some hard cases. As shown in the second row of Fig. 8, when the handbag is completely shadowed, the original red color can not be fully recovered.

5 Conclusion

This paper proposes a mask-guided residual learning network for joint single-image shadow detection and removal. We design a compact and efficient shadow detection module to generate shadow masks and feed them into our shadow removal module. To transfer context features between shadow detection and shadow removal, we reassemble and refine the features to generate shadow removal context features, which are further used to learn residual RGB maps to compensate for the input shadow map. Meanwhile, the predicted shadow mask serves as the guidance weight for fusing the residual and original RGB map, which helps to generate better transition effects in the penumbra regions. Extensive experiments demonstrate that our proposed network achieves state-of-the-art shadow detection and removal performance on three widely used benchmark datasets ISTD, ISTD+, and SRD.

Acknowledgements This work was supported by the National Natural Science Foundation of China (62076029), Guangdong Science and Technology Department (2017A030313362), Guangdong Key Lab of AI and Multi-modal Data Processing (2020KSYS007). and internal funds of the United International College (R202012, R201802, R5201904, UICR0400025-21).

References

1. Chen, Z., Zhu, L., Wan, L., Wang, S., Feng, W., Heng, P.A.: A multi-task mean teacher for semi-supervised shadow detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5611–5620 (2020) 1, 2.1
2. 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 (ICCV). pp. 4743–4752 (2021) 2.2, 4.4, 2, 4, 4.4
3. 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. pp. 10680–10687 (2020) 1, 5, 4.4, 2, 4, 6, 4.4
4. Ding, B., Long, C., Zhang, L., Xiao, C.: Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019) 1, 1
5. Finlayson, G.D., Drew, M.S., Lu, C.: Entropy minimization for shadow removal. International Journal of Computer Vision **85**(1), 35–57 (2009) 1, 2.1
6. 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), 59–68 (2006). https://doi.org/10.1109/TPAMI.2006.18 1, 2.1
7. 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 (CVPR). pp. 10571–10580 (2021) 1, 2.2, 5, 4.4, 4.4, 2, 3, 4
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of International Conference on Neural Information Processing Systems (NeurIPS). pp. 2672–2680 (2014) 2.1
9. Gryka, M., Terry, M., Brostow, G.J.: Learning to remove soft shadows. ACM Transactions on Graphics (TOG) **34**(5), 1–15 (2015) 2.2
10. Guo, R., Dai, Q., Hoiem, D.: Paired regions for shadow detection and removal. IEEE transactions on pattern analysis and machine intelligence **35**(12), 2956–2967 (2012) 2.1, 5, 4.4, 4.4, 2, 3, 4
11. 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), 2795–2808 (2019) 1, 2.2, 4.3, 5, 4.4, 1, 2, 4, 6, 4.4
12. 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 (ICCV). pp. 2472–2481 (2019) 1, 2.2, 4.1, 4.4, 4.4, 2, 3
13. Inoue, N., Yamasaki, T.: Learning from synthetic shadows for shadow detection and removal. IEEE Transactions on Circuits and Systems for Video Technology **31**(11), 4187–4197 (2021). https://doi.org/10.1109/TCSVT.2020.3047977 1
14. Jie, L., Zhang, H.: A fast and efficient network for single image shadow detection. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2634–2638 (2022) 1, 2.1
15. Jie, L., Zhang, H.: Rmlanet: Random multi-level attention network for shadow detection. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2022) 1, 2.1

16. 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 (ICCV). pp. 5027–5036 (2021) 4.1
17. Khan, S.H., Bennamoun, M., Sohel, F., Togneri, R.: Automatic feature learning for robust shadow detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1939–1946 (2014) 2.1
18. Lalonde, J.F., Efros, A.A., Narasimhan, S.G.: Detecting ground shadows in outdoor consumer photographs. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 322–335 (2010) 1, 2.1
19. Le, H., Samaras, D.: Shadow removal via shadow image decomposition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8578–8587 (2019) 1, 4.1, 4.1, 4.4, 3
20. Le, H., Samaras, D.: From shadow segmentation to shadow removal. In: European Conference on Computer Vision. pp. 264–281. Springer (2020) 4.4, 3
21. Nguyen, V., Vicente, T.F.Y., Zhao, M., Hoai, M., Samaras, D.: Shadow detection with conditional generative adversarial networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4510–4518 (2017) 1, 2.1
22. Qu, L., Tian, J., He, S., Tang, Y., Lau, R.W.H.: Deshadownet: A multi-context embedding deep network for shadow removal. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4067–4075 (2017) 1, 2.2, 4.1, 4.4, 3, 4, 6
23. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2015) 2.1
24. Shen, L., Chua, T.W., Leman, K.: Shadow optimization from structured deep edge detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2067–2074 (2015) 2.1
25. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems **28** (2015) 3.3
26. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML). pp. 6105–6114 (2019) 3.1
27. 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), 682–695 (2018) 2.1
28. 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 (CVPR). pp. 1788–1797 (2018) 1, 2.2, 4.1, 5, 1, 2
29. 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 (CVPR). pp. 1788–1797 (2018) 4.3, 5, 4.4, 4.4, 1, 2, 3
30. Xiao, C., She, R., Xiao, D., , Ma, K.L.: Fast shadow removal using adaptive multiscale illumination transfer. Computer Graphics Forum **32**, 6105–6114 (2019) 1
31. Xiaowei, H., Zhu, L., Fu, C.W., Qin, J., Heng, P.A.: Direction-aware spatial context features for shadow detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7454–7462 (2018) 1, 2.1

32. Yang, Q., Tan, K.H., Ahuja, N.: Shadow removal using bilateral filtering. IEEE Transactions on Image processing **21**(10), 4361–4368 (2012) 2.2
33. Zhang, L., Zhang, Q., Xiao, C.: Shadow remover: Image shadow removal based on illumination recovering optimization. IEEE Transactions on Image Processing **24**(11), 4623–4636 (2015) 2.2
34. Zhang, L., Zhang, Q., Xiao, C.: Shadow remover: Image shadow removal based on illumination recovering optimization. IEEE Transactions on Image Processing **24**(11), 4623–4636 (2015) 4.4, 4.4, 2, 3
35. Zheng, Q., Qiao, X., Cao, Y., Lau, R.W.: Distraction-aware shadow detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5167–5176 (2019) 1, 2.1, 4.3, 1
36. Zhu, J., Samuel, K.G., Masood, S.Z., Tappen, M.F.: Learning to recognize shadows in monochromatic natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 223–230 (2010) 2.1
37. Zhu, L., Deng, Z., Hu, X., Fu, C.W., Xu, X., Qin, J., Heng, P.A.: Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 121–136 (2018) 1, 2.1, 4.3, 1