Style-Guided Shadow Removal

Jin Wan[1], Hui Yin[1,✉], Zhenyao Wu[2], Xinyi Wu[2],
Yanting Liu[3], and Song Wang[2,✉]

[1] Beijing key lab of traffic data analysis and mining, Beijing Jiaotong University
[2] Department of Computer Science and Engineering, University of South Carolina
[3] Key Laboratory of Beijing for Railway Engineering, Beijing Jiaotong University
{jinwan, hyin}@bjtu.edu.cn, {zhenyao, xinyiw}@email.sc.edu,
19112024@bjtu.edu.cn, songwang@cec.sc.edu

Abstract. Shadow removal is an important topic in image restoration, and it can benefit many computer vision tasks. State-of-the-art shadow-removal methods typically employ deep learning by minimizing a pixel-level difference between the de-shadowed region and their corresponding (pseudo) shadow-free version. After shadow removal, the shadow and non-shadow regions may exhibit inconsistent appearance, leading to a visually disharmonious image. To address this problem, we propose a style-guided shadow removal network (SG-ShadowNet) for better image-style consistency after shadow removal. In SG-ShadowNet, we first learn the style representation of the non-shadow region via a simple region style estimator. Then we propose a novel effective normalization strategy with the region-level style to adjust the coarsely re-covered shadow region to be more harmonized with the rest of the image. Extensive experiments show that our proposed SG-ShadowNet outperforms all the existing competitive models and achieves a new state-of-the-art performance on ISTD+, SRD, and Video Shadow Removal benchmark datasets. Code is available at: https://github.com/jinwan1994/SG-ShadowNet.

Keywords: Shadow Removal, Region Style, Normalization

1 Introduction

Shadows are widespread in real-world images. In many computer-vision applications, shadows can be regarded as a kind of image degradation that undermines the information to be conveyed by the images and usually increases the difficulty of the downstream tasks [8,45,9,34,3,14,44]. In response to this problem, many shadow-removal approaches [12,10,32,40,15,21,22,23,7,29] have been developed in recent years, aiming to restore images to shadow-free ones. Shadow removal is still considered to be a very challenging problem due to various and complex shadow formation environments [1,26,46,10].

Shadow removal has been well studied from different perspectives, such as feature extraction [32,5,4], multi-task learning [40,15], image generation [49,28,29], image decomposition [22,23], and auto-exposure fusion [7]. The shadow-removal

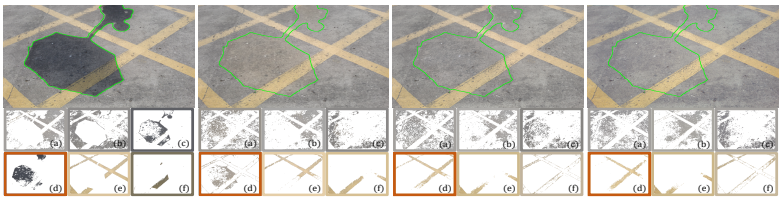

Fig. 1. De-shadowed images produced by the existing method of Fu *et al.* [7] and our proposed method. The color clustering results of each image are shown below it, by setting the number of clusters to 6, (a)∼(f). Compared with the clusters from Fu *et al.*, ours are more consistent with those from the shadow-free image, without separating out the shadow region, as shown by the cluster (d).

performance has been significantly improved in recent years by employing various advanced deep neural networks [22,5,23,29,7]. However, most of existing methods try to minimize certain pixel-level differences between the de-shadowed region and their corresponding (pseudo) shadow-free version, without explicitly considering the style consistency of de-shadowed and non-shadow regions. As a results, the image appearance of the de-shadowed and non-shadow regions may be inharmonious after shadow removal. An example is shown in Fig. 1, where the shadow region, after shadow removal using an existing method [7], is still visually distinguishable from the rest of image. In Fig. 1, we also use the color-based clustering to show the harmony of the de-shadowed results since color is an important cue in describing the image style [41] and style consistency [25]. We can see that the six clusters obtained from the image de-shadowed by the existing method are more aligned with those obtained from the original image with shadow, while the six clusters obtained from the image de-shadowed by our proposed method are more aligned with those obtained from the corresponding ground-truth shadow-free image.

In this paper, we reformulate shadow removal as an intra-image style transfer problem by explicitly considering the style consistency between shadow regions and non-shadow regions after shadow removal. Based on this formulation, we propose a new style-guided shadow removal network, namely SG-ShadowNet, which consists of a coarse deshadow network (CDNet) and a style-guided re-deshadow network (SRNet) by taking the whole model scale into account. The former employs a simple U-net structure to obtain a coarse de-shadowed result. The latter estimates the style representation of non-shadow regions and then uses it to help further refine the shadow removal, which is achieved by a new learnable spatially region-aware prototypical normalization (SRPNorm) layer for aligning the pixel-wise mean and variance between the de-shadowed and non-shadow regions. SRNet can also perform shadow removal without CDNet and achieve comparable shadow removal performance. To evaluate the proposed SG-ShadowNet, we conduct extensive experiments on the ISTD+ and SRD datasets and assess the generalization ability of the proposed method on the Video Shadow Removal dataset. In summary, the contributions of this work are as follows:

- To the best of our knowledge, this paper is the first work to study the problem of shadow removal from the perspective of *intra-image style transfer* and tackle it by preserving the whole image harmonization through the region style guidance.
- We propose a novel SG-ShadowNet, with a newly designed SRPNorm layer, to remove shadows while maintaining the style consistency between de-shadowed and non-shadow regions.
- The proposed SG-ShadowNet achieves new state-of-the-art performances on three public datasets, and it also exhibits strong generalization ability with fewer parameters.

2 Related Work

Before the deep learning era, shadow removal is usually achieved by extracting hand-crafted features and leveraging physical models of shadows [35,12,33] with limited performance. In this section, we mainly review state-of-the-art deep-learning approaches for shadow removal. We also go over existing works on the usage of normalization layers.

Shadow removal. Using deep learning, shadows can be removed by learning a complex mapping between shadow images/regions and the corresponding shadow-free images/regions with large-scale annotated training datasets [22,32].

The existing works formulate the shadow removal using different models, resulting in different algorithms. 1) From the feature extraction perspective, Qu *et al.* [32] proposed DeshadowNet to extract multi-context information and predict a shadow matte layer for removing shadows. Cun *et al.* [5] designed a dual hierarchically aggregation network (DHAN) to eliminate boundary artifacts by aggregating the dilated multi-context features and attentions. Most recently, Chen *et al.* [4] presented to remove shadows based on patch-level feature matching and transferring which cannot guarantee the global harmony of the whole image. 2) From the multi-task learning perspective, Wang *et al.* [40] employed a stacked conditional GAN to combine shadow detection and removal. Hu *et al.* [15] proposed to utilize direction-aware context to further improve the ability of shadow detection and removal. 3) From the image decomposition perspective, Le *et al.* [22], [23] and [24] employed a physical shadow illumination model to decompose the shadow images into different learnable parameters for generating shadow-free images, which implicitly considers image harmonization very roughly without adequately exploring the underlying relationship between shadow and non-shadow regions. 4) From the image generation perspective, Mask-shadowGAN [16] and LG-ShadowNet [28] leveraged GAN-based models to perform unsupervised shadow removal by learning a map between shadow domain and non-shadow domain. Recently, Liu *et al.* [29] developed a shadow generation model with shadow mask to construct pseudo shadows and shadow-free image pairs for weakly-supervised shadow removal. 5) From the auto-exposure fusion perspective, Fu *et al.* [7] proposed an auto-exposure fusion network, which utilizes shadow-aware fusion network to adaptively fuse the estimated multiple over-exposure images to generate the shadow-free image.

As mentioned earlier, these methods do not explicitly consider the style consistency between shadow and non-shadow regions after shadow removal, and therefore may lead to an image disharmony. In this paper, we focus on addressing this problem by developing a new SG-ShadowNet with explicit consideration of the style consistency in the same image.

Normalization layers. Normalization of the training data can improve the similarity of data distribution and facilitate the network optimization [36]. Normalizing the intermediate representation of deep networks can also improve the network prediction performance, which has led to many studies on the use of normalization layers in deep learning. Two kinds of normalization have been used for network layers: unconditional normalization [18,37,2,43,36] and conditional normalization [39,6,17,31,47,27,25]. The former does not use external data to provide affine parameters, which is irrelevant to our work. The latter normalizes the mean and deviation of feature maps and then uses external data to learn the affine transformation parameters to denormalize the feature map. Huang *et al.* [17] proposed an adaptive instance normalization (AdaIN) for real-time image stylization, which uses a pre-trained VGG network to extract the style representation of other images. This is not applicable to our task, because the style representation of the non-shadow region cannot be extracted by a pre-trained network. Ling *et al.* [25] proposed a region-aware adaptive instance normalization (RAIN) for image harmonization where the foreground feature is normalized with the channel-wise affine parameters predicted by the background feature in the same intermediate feature map. This inspires us to take the non-shadow region as external data to generate the style representation for adjusting the shadow region. Different from [25], 1) we adopt a coarse-to-fine network to mitigate the difficulty of style transfer directly from shadow to non-shadow version, 2) we employ a region style estimator to accurately learn the style representation of the non-shadow region which later provides guidance for shadow removal, and 3) we design a spatially region-aware normalization layer, which can estimate pixel-wise affine parameters to capture the spatial-variant property of shadows.

3 Methodology

3.1 Problem formulation

In this work, we propose to reformulate the shadow removal as an intra-image style transfer problem, *i.e.*, the style representation learned from the non-shadow region is applied to the shadow removal of the shadow region, so that the de-shadowed region holds the similar style, such as color and lighting, as the non-shadow region. Specifically, we render the shadow region of the original shadow image I^{S} by using the style of the non-shadow region of the same image to achieve a de-shadowed image $\hat{I}$ with consistent visual styles. It can be formulated as:

$$\hat{I} = \psi(I^{\mathrm{S}}, M|P), \tag{1}$$

where $\psi(\cdot,\cdot)$ represents a style transfer function, P denotes the non-shadow prototype, and M indicates the shadow mask (region) in I^{S}. Meanwhile, consid-

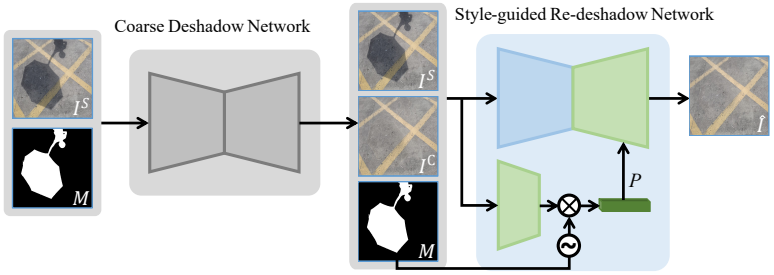

Fig. 2. An illustration of the proposed SG-ShadowNet. It sets the non-shadow prototype P as prior information to adjust the shadow region of the coarsely de-shadowed results I^{C}, resulting in a visual consistent shadow-free image $\hat{I}$.

ering the difficulty in stylizing directly from shadow to non-shadow version, we achieve the image stylization on a coarsely de-shadowed result I^{C}, by reformulating Eq. (1) as

$$\hat{I} = \psi(I^{\mathrm{C}}, M|P) = \psi(\mathcal{G}(I^{\mathrm{S}}, M), M|P), \tag{2}$$

where $\mathcal{G}(\cdot,\cdot)$ represents the coarse deshadow network which takes the original shadow image I^{S} and the corresponding shadow mask M as inputs.

Fig. 2 shows the overall framework of the proposed shadow-removal network, which consists of two stages. In the first stage, a coarse deshadow network (CDNet) is utilized to obtain I^{C} for alleviating the difficulty of style transfer. We use the U-net structure in [13] as its backbone and remove all skip connections and half of the filters to reduce computational complexity. In the second stage, we propose a style-guided re-deshadow network (SRNet) in Section 3.2, which uses the estimated style representation of the non-shadow region to adaptively adjust the style of the shadow region in the same image. In addition, given that shadows present spatial-variant property, *i.e.*, the color and illumination distortion across shadow region are variant, we propose a novel spatially region-aware prototypical normalization (SRPNorm) in Section 3.3 to adjust the coarsely recovered shadow region in I^{C} to be more harmonious with the rest of the image. Note that we can also perform SRNet without CDNet by using I^{S} and M as inputs for shadow removal, and the results are shown in Table 4.

3.2 Style-guided re-deshadow network

In this section, we elaborate on the style-guided re-deshadow network (SRNet), which is composed of a light-weight region style estimator and a re-deshadow network, as shown in Fig. 3. To accurately obtain the style representation of the realistic non-shadow region, we composite the coarsely predicted result I^{C} and the shadow image I^{S} using shadow mask M as one of the inputs of the region style estimator and the re-deshadow network, which can be calculated as

$$I^{\mathrm{in}} = I^{\mathrm{C}} \otimes M + I^{\mathrm{S}} \otimes \bar{M} \tag{3}$$

where $\otimes$ is the Hadamard product, and $\bar{M} = 1 - M$ represents the mask of non-shadow region. The region style estimator takes I^{in} and $\bar{M}$ as inputs to

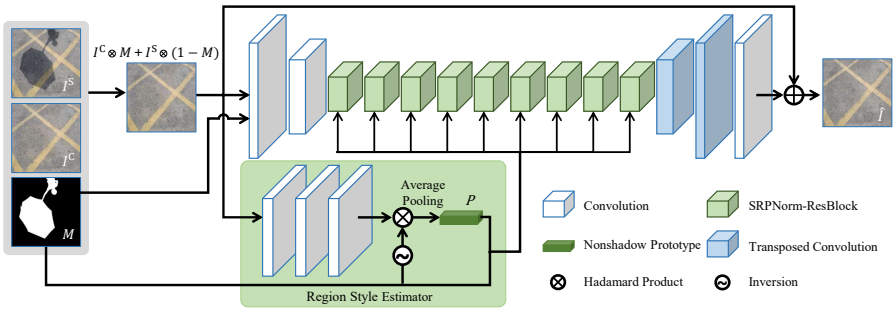

Fig. 3. An illustration of the proposed style-guided re-deshadow network.

get non-shadow prototype P. The re-deshadow network takes I^{in}, M and P as inputs to generate the final shadow-free image $\hat{I}$.

Region style estimator. To incorporate the style information of the non-shadow region, a newly designed region style estimator is proposed to learn the non-shadow prototype P for the re-deshadow network, which consists of three convolution layers and a global pooling layer. Note that: **1)** To avoid interference of non-shadow and shadow regions in the same image, we restrict the receptive field of the estimator by using 1×1 kernels for all the convolution layers. **2)** To obtain the accurate style representation of the non-shadow region, we perform the Hadamard product on the $\bar{M}$ and the unpooled features to ensure that the output prototype P is only related to the non-shadow region. The details of the estimator are depicted in Fig. 3 (bottom).

Re-deshadow network. As shown in Fig. 3 (top), the architecture of the re-deshadow network follows the U-Net [28] and includes 9 residual blocks in the middle. One unique trait of the re-deshadow network is that we embed the proposed spatial region prototypical normalization (SRPNorm) in each residual block, called SRPNorm-ResBlock. The SRPNorm-ResBlock consists of two convolutions and two SRPNorm modules, as shown in Fig. 4.

3.3 Spatially Region-aware Prototypical Normalization (SRPNorm)

In each SRPNorm-ResBlock, the proposed SRPNorm module utilizes the non-shadow prototype $P \in \mathbb{R}^{1\times1\times C}$ and the resized shadow mask $M \in \mathbb{R}^{H\times W\times1}$ as conditional inputs to perform affine transformation on the input feature map $F^{\text{in}} \in \mathbb{R}^{H\times W\times C}$, where H, W, C are the height, width, and channel number of the feature maps.

There are two options for the affine parameters here. One is to learn channel-wise affine parameters, *i.e.*, the pixels on each channel are affinely transformed with the same scale and bias. The other is to learn pixel-wise affine parameters, *i.e.*, each pixel has its own individually adapted scale and bias for affine transformation. Considering the spatial-variant property of the shadows, we adopt the latter to perform a pixel-wise affine transformation on shadow regions.

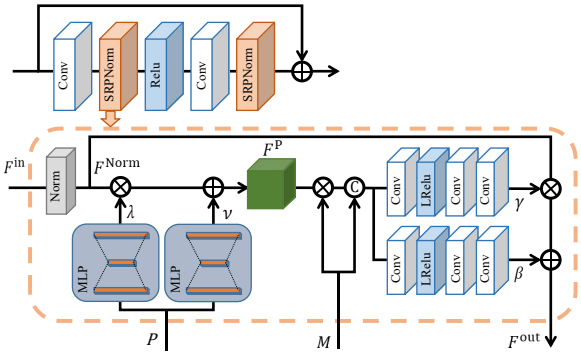

Fig. 4. An illustration of the proposed SRPNorm-ResBlock.

Specifically, we first perform a region normalization on F^{in} to obtain the normalized features F^{Norm}, which can be calculated by

$$
\begin{aligned}
F^{\text{Norm}}_{h,w,c} &= \frac{F^{\text{in}}_{h,w,c} \otimes M_{h,w} - \mu_{M,c}}{\sqrt{\delta^2_{M,c} + \epsilon}} + \frac{F^{\text{in}}_{h,w,c} \otimes \bar{M}_{h,w} - \mu_{\bar{M},c}}{\sqrt{\delta^2_{\bar{M},c} + \epsilon}}, \\
\mu_{M,c} &= \frac{1}{\sum\limits_{h,w} M_{h,w}} \sum_{h,w} (F^{\text{in}}_{h,w,c} \otimes M_{h,w}), \\
\mu_{\bar{M},c} &= \frac{1}{\sum\limits_{h,w} \bar{M}_{h,w}} \sum_{h,w} (F^{\text{in}}_{h,w,c} \otimes \bar{M}_{h,w}), \\
\delta_{M,c} &= \sqrt{\frac{1}{\sum\limits_{h,w} M_{h,w}} \sum_{h,w} \left(M_{h,w} \otimes (F^{\text{in}}_{h,w,c} - \mu_{M,c})^2 \right)}, \\
\delta_{\bar{M},c} &= \sqrt{\frac{1}{\sum\limits_{h,w} \bar{M}_{h,w}} \sum_{h,w} \left(\bar{M}_{h,w} \otimes (F^{\text{in}}_{h,w,c} - \mu_{\bar{M},c})^2 \right)},
\end{aligned}
\tag{4}
$$

where μ_M, δ_M and $\mu_{\bar{M}}$, $\delta_{\bar{M}}$ are channel-wise average and standard deviation of shadow and non-shadow regions in F^{in}. ϵ is set to 1e-5.

Since shadows present spatial-variant property, each pixel in the shadow region is subject to different affine parameters. We further utilize P to modulate F^{Norm} and compute the spatial prior information F^{P} of the non-shadow region. This is achieved by sending P to two MLPs and generating two channel-wise modulated parameters $\lambda_c(P)$ and $\nu_c(P)$ for F^{Norm}. F^{P} can be expressed by

$$
F^{\text{P}}_{h,w,c} = \lambda_c(P) \otimes F^{\text{Norm}}_{h,w,c} + \nu_c(P). \tag{5}
$$

Then we learn the pixel-wise affine parameters $\gamma_{h,w,c}(F^{\text{P}}, M)$ and $\beta_{h,w,c}(F^{\text{P}}, M)$ by using three convolution layers and taking F^{P} and M as input. Finally, we perform affine processing on the normalized features F^{Norm} based on the scale (γ) and bias (β) learned from non-shadow regions. The output of SRPNorm is defined as:

$$
F^{\text{out}}_{h,w,c} = \gamma_{h,w,c}(F^{\text{P}}, M) \otimes F^{\text{Norm}}_{h,w,c} + \beta_{h,w,c}(F^{\text{P}}, M). \tag{6}
$$

3.4 Loss function

For the coarse deshadow network, the pixel-level reconstruction loss L_{r1} is used to optimize the distance between the ground truth shadow-free image I^{SF} and the coarsely de-shadowed image I^{C}:

$$L_{\mathrm{r1}} = \|I^{\mathrm{C}} - I^{\mathrm{SF}}\|_1. \tag{7}$$

Moreover, we use area loss L_{a1} to strengthen the constraint of the shadow region as in previous work [29]. Formally, the area loss is defined as

$$L_{\mathrm{a1}} = \|\phi(M) \otimes I^{\mathrm{C}} - \phi(M) \otimes I^{\mathrm{SF}}\|_1, \tag{8}$$

where $\phi(\cdot)$ denotes the image dilation function with a kernel size of 50 and $\otimes$ is the Hadamard product.

For the style-guided re-deshadow network (SRNet), the output of SRNet is the final de-shadowed result $\hat{I}$. We also calculate the reconstruction loss L_{r2} and the area loss L_{a2} between $\hat{I}$ and I^{SF} as

$$\begin{aligned} L_{\mathrm{r2}} &= \|\hat{I} - I^{\mathrm{SF}}\|_1, \\ L_{\mathrm{a2}} &= \|\phi(M) \otimes \hat{I} - \phi(M) \otimes I^{\mathrm{SF}}\|_1. \end{aligned} \tag{9}$$

To ensure the spatial consistency of $\hat{I}$, we apply the spatial consistency loss [11]:

$$L_{\mathrm{s}} = \frac{1}{K} \sum_{i=1}^{K} \sum_{j \in \Omega(i)} \left(|(Y_i, Y_j)| - |(V_i, V_j)|\right)^2, \tag{10}$$

where K denotes the number of local areas, $\Omega(i)$ represents four adjacent areas centered at area i, and Y and V are the average intensity values of the local areas of $\hat{I}$ and I^{SF}, respectively. Finally, we define the total loss function $\mathcal{L}$ as

$$\mathcal{L} = L_{\mathrm{a1}} + L_{\mathrm{r1}} + L_{\mathrm{a2}} + L_{\mathrm{r2}} + \zeta L_{\mathrm{s}}, \tag{11}$$

where ζ denotes the weight term of spatial consistency loss and is empirically set to 10. In our experiments, we set the weights of reconstruction losses and area losses to 1 by following [29].

4 Experiments

4.1 Experimental setup

Datasets. We train and evaluate the proposed method on the ISTD+ [22] and SRD [32] datasets and verify the generalization of our model on the Video Shadow Removal dataset [23]. **1)** The ISTD+ dataset has 1,870 triplets of shadow, shadow-free, and shadow mask images, where 1,330 triplets are used for training and the remaining 540 triplets are used for testing. We use the provided ground-truth shadow mask in the training phase, while for the test, the corresponding shadow mask of the test shadow image is calculated by a pre-trained BDRAR shadow detector [50] that trained on the SBU [38] and ISTD+

datasets. The Balanced Error Rate of the model in the ISTD+ testing set is 2.4. **2)** SRD dataset contains 2,680 training pairs of shadow and shadow-free images and 408 testing pairs. Same as [7], we utilize Otsu's algorithm [30] to extract the shadow mask from the difference between shadow-free and shadow images during training, and we exploit the shadow masks detected by DHAN [5] for testing. **3)** Video Shadow Removal dataset consists of 8 videos captured in the static scene, *i.e.*, there are no moving objects in each video. As [23], we employ a threshold of 40 to get the moving shadow mask for evaluation which divides shadow and non-shadow pixels according to intensity difference. In addition, we utilize a pre-trained BDRAR [50] to generate shadow masks for testing.

Evaluation metrics. We employ the root mean square error (RMSE [4]) in the LAB color space, and we adopt the learned perceptual image patch similarity (LPIPS) [48] to evaluate the perceptual quality of the de-shadowed results.

Implementation details. We implement the proposed network using PyTorch with a single NVIDIA GeForce GTX 2080Ti GPU card. In our experiments, the coarse deshadow network (CDNet) and style-guided re-deshadow network (SRNet) are jointly trained to obtain the final shadow-free image. For data augmentation, we exploit random flipping and random cropping with a crop size of 400×400. During training, our model is optimized by Adam [20] with the first and the second momentum being set to 0.50 and 0.99, respectively, and the batch size is set to 1. The basic learning rate is set to 2×10^{-4} and halved every 50 epochs with 200 epochs in total.

4.2 Comparison with state-of-the-arts

We compare our SG-ShadowNet with 14 state-of-the-art shadow removal algorithms, including unsupervised methods of Mask-ShadowGAN [40], LG-ShadowNet [28], and DC-shadowNet [19], weakly supervised methods of Gong & Cosker [10], Param+M+D-Net [23], and G2R-ShadowNet [29], and fully supervised methods of ST-CGAN [40], DeshadowNet [32], SP+M-Net [22], DSC [15], DHAN [5], CANet [4], Fu *et al.* [7], and SP+M+I-Net [24]. For a fair comparison, all results are taken from their original papers or generated by their official code.

Quantitative evaluation. Tables 1 and 2 quantitatively show the test results of different shadow-removal methods on the ISTD+ and SRD datasets. Compared with unsupervised, weakly-supervised and supervised methods, SG-ShadowNet performs the best on both shadow regions and the whole image. Specifically, on the ISTD+ dataset, our method outperforms Fu *et al.* [7] by decreasing the RMSE of the shadow region and entire image by 9.2% and 19.0%, respectively. Our method also outperforms SP+M+I-Net [24] by a small margin in shadow regions. Meanwhile, our method obtains a lower LPIPS score than other existing methods, which verifies that our de-shadowed results show a more consistent style with the shadow-free images, and further proves the effectiveness of the style guidance strategy. On the SRD dataset, the public shadow masks generated by [5] are employed for evaluation. Our method outperforms the fully-supervised

[4] The RMSE is actually calculated by the mean absolute error (MAE) as [22].

Table 1. Shadow removal results of the proposed method compared to state-of-the-art shadow removal methods on ISTD+ [22]. RMSE and LPIPS are the lower the better.

Scheme	Method	**Shadow** RMSE↓	**Non-Shadow** RMSE↓	**All** RMSE↓	**All** LPIPS↓
Un-supervised	Mask-ShadowGAN [40]	9.9	3.8	4.8	0.095
	LG-ShadowNet [28]	9.7	3.4	4.4	0.103
	DC-ShadowNet [19]	10.4	3.6	4.7	0.170
Weakly-supervised	Gong & Cosker [10]	13.3	**2.6**	4.3	0.086
	Param+M+D-Net [23]	9.7	2.9	4.1	0.086
	G2R-ShadowNet [29]	8.8	2.9	3.9	0.096
Fully-supervised	ST-CGAN [40]	13.4	7.9	8.6	0.150
	SP+M-Net [22]	7.9	2.8	3.6	0.085
	Fu *et al.* [7]	6.5	3.8	4.2	0.106
	SP+M+I-Net [24]	6.0	3.1	3.6	0.092
	SG-ShadowNet (Ours)	**5.9**	2.9	**3.4**	**0.070**

Table 2. Shadow removal results of the proposed method compared to state-of-the-art shadow removal methods on SRD [32]. '*' indicates that the result is directly cited from the original paper.

Method	**Shadow** RMSE↓	**Non-Shadow** RMSE↓	**All** RMSE↓	**All** LPIPS↓
DeshadowNet [32]	11.78	4.84	6.64	0.165
DSC [15]	10.89	4.99	6.23	0.147
DHAN [5]	8.94	4.80	5.67	0.104
Fu *et al.* [7]	8.56	5.75	6.51	0.153
CANet* [4]	7.82	5.88	5.98	-
DC-ShadowNet [19]	8.26	3.68	4.94	0.167
SG-ShadowNet (Ours)	**7.53**	**2.97**	**4.23**	**0.099**

Fu *et al.* [7] and unsupervised DC-ShadowNet [19] in shadow regions, reducing RMSE by 12.0% and 8.8%, respectively. It also decreases the RMSE from 7.82 to 7.53, compared to CANet [4]. Moreover, it can be seen from Table 3 that SG-ShadowNet has only 6.2M parameters, which is less than 4.4% and 3.2% of the shadow removal network parameters in [7] and [24], respectively.

Qualitative evaluation. Fig. 5 provides the visual comparisons of shadow-removal results produced by different methods. It can be easily observed that the existing methods suffer inconsistent appearance between the shadow and non-shadow regions after shadow removal. On the contrary, our SG-ShadowNet can produce visually more harmonious de-shadowed images.

4.3 Ablation study

Effectiveness of SRPNorm. To investigate the effect of SRPNorm, we conduct the following ablations: 1) replacing all SRPNorms with other normalization

Table 3. Number of parameters and flops of SG-ShadowNet and other comparison methods, with input size of 256 × 256.

Method	#Params.	Flops
SP+M-Net [22]	141.2 M	39.8 G
G2R-ShadowNet [29]	22.8 M	113.9 G
Fu *et al.* [7]	142.2 M	104.8 G
SP+M+I-Net [24]	195.6 M	58.0 G
SG-ShadowNet(Ours)	6.2 M	39.7 G

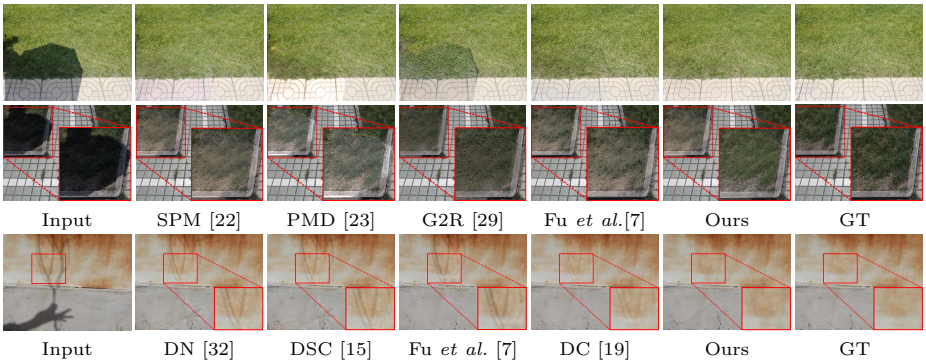

Fig. 5. Visualisation comparisons on the ISTD+ [22] (top two rows) and SRD [32] (bottom row) datasets.

layers (*i.e.*, BN [18], IN [37], RN [47], and RAIN [25]); 2) using SRPNorm to provide different style-guided affine parameters (channel- and pixel- wise) for the normalization layer; and 3) adding k SRPNorm into the innermost ResBlocks of the SRNet (SRPNorm-k) with different k values. The results are reported in the Table 4.

We first apply the classic BN and IN as the normalization layer of the network. We can see that their performances are limited since they do not use additional conditions to de-normalize the features of shadow regions. Then we deploy RN and RAIN into our network. Although RN can use external conditions to de-normalize shadow and non-shadow regions separately, it strictly distinguishes the features of shadow and non-shadow regions, preventing the information of non-shadow regions from spreading to shadow regions. The RAIN, an enhanced version of RN, utilize the non-shadow region information within the feature maps to de-normalize features of shadow regions, where the information between shadow and non-shadow regions would interfere with each other in the middle feature maps, so it cannot accurately reflect the style of non-shadow regions. Moreover, RAIN adopts channel-wise normalization, *i.e.*, the same mean and variance are used to de-normalize the features of shadow regions. This obviously does not consider the spatial-variant property of shadows, which makes it unable to generalize well to the shadow removal task. As shown in Fig. 6, without

Table 4. Ablation study of the proposed SRPNorm on ISTD+ [22]

Model	**Shadow** RMSE↓	**Non-Shadow** RMSE↓	**All** RMSE↓	**All** LPIPS↓
BN [18]	7.5	3.8	4.0	0.099
IN [42]	7.3	2.9	3.7	0.076
RN [47]	6.7	3.0	3.6	0.073
RAIN [25]	6.6	2.9	3.5	0.074
SRPNorm w/o S	6.3	2.9	3.4	0.072
SRPNorm w/o P	6.7	2.9	3.5	0.073
SRPNorm w/o M	6.2	2.9	3.4	0.072
SRPNorm-3	6.3	2.9	3.5	0.071
SRPNorm-5	6.1	2.9	3.4	0.071
SRPNorm-7	6.0	2.9	3.4	0.070
SRPNorm-All	5.9	2.9	3.4	0.070

Fig. 6. Visualisation comparisons of different normalization methods.

the style guidance of the non-shadow region, the de-shadowed results of IN and RN show obvious color and lighting distinction with the original non-shadowed part of this image. In addition, the color of de-shadowed results from RAIN is not consistent with the neighbor regions since it's hard to accurately extract the non-shadow region style. Obviously, our method achieves more visual consistent results than other normalization-based methods. The numerical performance on the ISTD+ dataset in Table 4 also verifies our observation.

Besides, SRPNorm can also provide channel-wise normalization (SRPNorm w/o S), *i.e.*, the results of Eq. (5) is the output of SRPNorm. It can be found in Table 4 that SRPNorm w/o S leads to a performance drop, which verifies the effectiveness of performing the pixel-wise (spatial) affine transformation on de-shadowed regions. Note that even if SRPNorm degenerates to a channel-wise normalization, it still outperforms the above-mentioned normalization methods, by benefiting from accurately extracting the style of the non-shadow region via the region style estimator. We also try to replace the non-shadow prototype in SRPNorm (SRPNorm w/o P) by only using the shadow mask as the prior information of SRPNorm. The decreased performance further verifies the effectiveness of the proposed style guidance from the non-shadow region. We then remove the shadow mask in Eq. (6) (SRPNorm w/o M) –we can see a slight drop in performance, which motivates us to focus on shadow regions during de-normalization. Finally, we try to insert different numbers of SRPNorms into the innermost ResBlocks of SRnet. It is obvious that shadow-removal performance

Table 5. Ablation study on the effectiveness of network architecture and loss functions on ISTD+ [22].

Model	**Shadow** RMSE↓	**Non-Shadow** RMSE↓	**All** RMSE↓	**All** LPIPS↓
w/o *CDNet*	6.5	2.9	3.5	0.076
w/o *SRNet*	7.0	2.9	3.5	0.072
w/o L_{a1}, L_{a2}	6.0	2.8	3.3	0.068
w/o L_s	6.2	3.0	3.5	0.073
SG-ShadowNet (Ours)	5.9	2.9	3.4	0.070

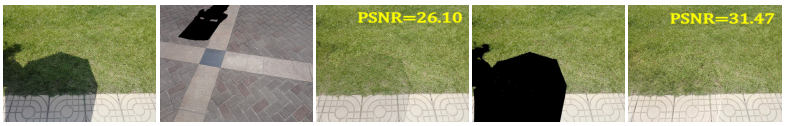

Fig. 7. Visualisation comparisons of the intra- vs. inter- image region style guidance capability for shadow removal.

is improved with the increase of the number of SRPNorm-ResBlocks (involving more layers for style transfer), which shows the superiority of SPRNorm.

Effectiveness of network architecture and loss function. We also provide ablation experiments to verify the contribution of the designed network architecture and loss function. From the first two rows of Table 5, we can see that the coarse deshadow network and style-guided re-deshadow network well complement each other – the shadow-removal performance drops by removing any one of these two networks. Note that the shadow removal performance on ISTD+ achieved by SRNet alone, *i.e.*, a one-stage processing, is comparable to Fu *et al.* [7], which is an impressive result. The remaining rows of Table 5 show the effectiveness of the area loss and the spatial consistency loss, without which RMSE of the shadow region increases by 0.1 and 0.3, respectively.

Intra- vs. inter- image style transfer. With the region style estimator, we are able to perform style transfer not only with the region style of the desired image (*Style-2*), but also with the region style of an irrelevant reference image (*Style-1*). From Fig. 7, it is obvious that *Result-1* by performing the style transfer with the latter is less harmonious and exists detailed style-related (*i.e.*, color) traces, which verifies the superiority of intra-image style transfer for shadow removal.

4.4 Generalization ability

To verify the generalization ability of our method, we compare it with several state-of-the-art methods, including SP+M-Net [22], Param+M+D-Net [23], Mask-ShadowGAN [16], LG-ShadowNet [28], G2R-ShadowNet [29], and DC-ShadowNet [19] on the Video Shadow Removal dataset [23]. All compared methods are pre-trained on ISTD+ [22] and tested directly on the video dataset.

Table 6. Shadow removal results on Video Shadow Removal dataset [23].

Method	RMSE↓	PSNR↓	SSIM↓
SP+M-Net [22]	22.2	-	-
Param+M+D-Net [23]	20.9	-	-
Mask-ShadowGAN [16]	19.6	19.47	0.850
LG-ShadowNet [28]	18.3	19.90	0.843
G2R-ShadowNet [29]	18.8	20.00	0.838
DC-ShadowNet [19]	18.9	19.92	0.848
SG-ShadowNet (Ours)	**16.5**	**21.65**	**0.852**

Fig. 8. Visual comparisons on the Video Shadow Removal dataset [23].

From Table 6, we see that our method performs best on all evaluation metrics, and outperforms the fully-supervised method SP+M-Net with RMSE decreased by 25.7% in the shaded region, and also outperforms the recent weakly supervised G2R-ShadowNet and unsupervised DC-ShadowNet. By using the style of non-shadow regions in an image as guidance for shadow removal, our proposed SG-ShadowNet exhibits better generalization ability in unknown environments, which also can be seen in the comparison of the qualitative results in Fig. 8.

5 Conclusion

In this paper, we proposed a style-guided shadow removal network (SG-ShadowNet) to achieve better image-style consistency after shadow removal. SG-ShadowNet can accurately learn the style representation of the non-shadow region using the regional style estimator, and employ the proposed spatially region-aware prototypical normalization (SRPNorm) to render the non-shadow region style to the shadow region on a coarsely de-shadowed image. Experimental results showed that the proposed SG-ShadowNet achieves the new state-of-the-art shadow removal performance on the ISTD+, SRD, and Video Shadow Removal datasets.

Acknowledgments: This work was supported by the Fundamental Research Funds for the Central Universities (2020YJS031), National Nature Science Foundation of China (51827813, 61472029, U1803264), and Research and Development Program of Beijing Municipal Education Commission (KJZD20191000402).

References

1. Arbel, E., Hel-Or, H.: Shadow removal using intensity surfaces and texture anchor points. IEEE Trans. Pattern Anal. Mach. Intell. **33**(6), 1202–1216 (2011)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Gool, L.V.: One-shot video object segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5320–5329 (2017)
4. Chen, Z., Long, C., Zhang, L., Xiao, C.: Canet: A context-aware network for shadow removal. In: Int. Conf. Comput. Vis. pp. 4743–4752 (October 2021)
5. Cun, X., Pun, C.M., Shi, C.: Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In: AAAI. pp. 10680–10687 (2020)
6. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. arXiv preprint arXiv:1610.07629 (2017)
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: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10571–10580 (June 2021)
8. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (June 2014)
9. Girshick, R.B.: Fast rcnn. In: Int. Conf. Comput. Vis. pp. 1440–1448 (2015)
10. Gong, H., Cosker, D.: Interactive shadow removal and ground truth for variable scene categories. In: Brit. Mach. Vis. Conf. (2014)
11. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S.T.W., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1777–1786 (2020)
12. Guo, R., Dai, Q., Hoiem, D.: Paired regions for shadow detection and removal. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2956–2967 (2012)
13. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1712–1722 (2019)
14. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: Int. Conf. Comput. Vis. (Oct 2017)
15. Hu, X., Fu, C.W., Zhu, L., Qin, J., Heng, P.A.: Direction-aware spatial context features for shadow detection and removal. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
16. Hu, X., Jiang, Y., Fu, C.W., Heng, P.A.: Mask-shadowgan: Learning to remove shadows from unpaired data. In: Int. Conf. Comput. Vis. (2019)
17. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Int. Conf. Comput. Vis. (2017)
18. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proc. Int. Conf. Mach. Learn. vol. 37, p. 448–456 (2015)
19. Jin, Y., Sharma, A., Tan, R.T.: Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In: Int. Conf. Comput. Vis. pp. 5027–5036 (2021)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2015)

21. Le, H., Goncalves, B., Samaras, D., Lynch, H.: Weakly labeling the antarctic: The penguin colony case. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 18–25 (2019)
22. Le, H., Samaras, D.: Shadow removal via shadow image decomposition. In: Int. Conf. Comput. Vis. (2019)
23. Le, H., Samaras, D.: From shadow segmentation to shadow removal. In: Eur. Conf. Comput. Vis. (2020)
24. Le, H., Samaras, D.: Physics-based shadow image decomposition for shadow removal. IEEE Trans. Pattern Anal. Mach. Intell. (2021)
25. Ling, J., Xue, H., Song, L., Xie, R., Gu, X.: Region-aware adaptive instance normalization for image harmonization. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9361–9370 (June 2021)
26. Liu, F., Gleicher, M.: Texture-consistent shadow removal. In: Eur. Conf. Comput. Vis. pp. 437–450 (2008)
27. Liu, H., Wan, Z., Huang, W., Song, Y., Han, X., Liao, J.: Pd-gan: Probabilistic diverse gan for image inpainting. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9371–9381 (June 2021)
28. Liu, Z., Yin, H., Mi, Y., Pu, M., Wang, S.: Shadow removal by a lightness-guided network with training on unpaired data. IEEE Trans. Image Process. **30**, 1853–1865 (2021)
29. Liu, Z., Yin, H., Wu, X., Wu, Z., Mi, Y., Wang, S.: From shadow generation to shadow removal. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
30. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. Syst. **9**(1), 62–66 (1979)
31. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2332–2341 (2019)
32. Qu, L., Tian, J., He, S., Tang, Y., Lau, R.W.: Deshadownet: A multi-context embedding deep network for shadow removal. In: IEEE Conf. Comput. Vis. Pattern Recog. (2017)
33. Shechtman, Eli, Li-Qian, Sunkavalli, Kalyan, Shi-Min, Wang, Jue: Appearance harmonization for single image shadow removal. European Association for Computer Graphics **35**(7), 189–197 (2016)
34. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 640–651 (2017)
35. Shor, Y., Lischinski, D.: The shadow meets the mask: Pyramid-based shadow removal. Comput. Graph. Forum **27**, 577–586 (04 2008)
36. Singh, S., Krishnan, S.: Filter response normalization layer: Eliminating batch dependence in the training of deep neural networks. In: IEEE Conf. Comput. Vis. Pattern Recog. (June 2020)
37. Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
38. Vicente, T.F.Y., Hou, L., Yu, C.P., Hoai, M., Samaras, D.: Large-scale training of shadow detectors with noisily-annotated shadow examples. In: Eur. Conf. Comput. Vis. pp. 816–832. Springer (2016)
39. de Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., Courville, A.C.: Modulating early visual processing by language. In: Adv. Neural Inform. Process. Syst. (2017)
40. Wang, J., Li, X., Yang, J.: Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018)

41. Wang, P., Li, Y., Vasconcelos, N.: Rethinking and improving the robustness of image style transfer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 124–133 (June 2021)
42. Wang, T., Hu, X., Wang, Q., Heng, P.A., Fu, C.W.: Instance shadow detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1880–1889 (2020)
43. Wu, Y., He, K.: Group normalization. In: Eur. Conf. Comput. Vis. (2018)
44. Wu, Z., Wu, X., Zhang, X., Wang, S., Ju, L.: Semantic stereo matching with pyramid cost volumes. In: Int. Conf. Comput. Vis. pp. 7483–7492 (2019)
45. Xie, S., Tu, Z.: Holistically-nested edge detection. Int. J. Comput. Vis. **125**, 3–18 (2015)
46. Yang, Q., Tan, K.H., Ahuja, N.: Shadow removal using bilateral filtering. IEEE Trans. Image Process. **21**(10), 4361–4368 (2012)
47. Yu, T., Guo, Z., Jin, X., Wu, S., Chen, Z., Li, W., Zhang, Z., Liu, S.: Region normalization for image inpainting. In: AAAI. pp. 12733–12740 (2020)
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 586–595 (2018)
49. Zhang, S., Liang, R., Wang, M.: Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. Computational Visual Media **5**(1), 105–115 (2019)
50. 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: Eur. Conf. Comput. Vis. pp. 121–136 (2018)