Towards Real-world Shadow Removal with a Shadow Simulation Method and a Two-stage Framework

Jianhao Gao
Wuhan University
johngao@whu.edu.cn

Quanlong Zheng*
OPPO Research Institute
xiaolong921001@gmail.com

Yandong Guo
OPPO Research Institute
yandong.guo@live.com

Abstract

Shadow removal is an important yet challenging restoration task. State-of-the-art shadow removal methods usually require paired datasets for training. Existing shadow removal datasets lack large-scale quantity and scene diversity. Hence, models trained on such datasets have poor generalization ability. This paper proposes a simple yet robust shadow simulation method to simulate shadow on the grayscale. The proposed shadow simulation method can be applied to arbitrary shadow-free images and masks to generate corresponding shadow images. With our shadow simulation method, we can generate a large-scale and diverse paired shadow removal dataset. Besides, we introduce a two-stage framework, Gray-to-Color Shadow Removal Network (G2C-DeshadowNet), for shadow removal. The first stage is a Grayscale Enhancement Network, which attempts to remove shadows on the grayscale. The second stage is a Colorization Network, which attempts to colorize the grayscale shadow-free image. Extensive experiments on ISTD+, SRD, and SBU datasets show that G2C-DeshadowNet outperforms state-of-the-art methods and has better generalization ability. We will release our code at https://github.com/jianhaogao/Shadow-Removal-with-Two-stage-Framework.

1. Introduction

Shadow is a common phenomenon in nature, formed by the light source being blocked by objects. It negatively affects vision tasks such as detection [4] and object tracking [26] since dark objects are easily confused with shadows [38]. Removing shadows is a meaningful and challenging task which has been long-term studied.

Traditional methods develop physical models to remove shadows, viewing shadow pixels as the combination of their illumination and reflection [8,9,13,14,25,26,33,37]. Hence, shadow-free pixels can be restored from the corresponding shadow pixels through an estimation of their illumination parameters. Yet, estimating the parameters requires user interactions [3, 37] and amounts of time, and thus cannot meet the need of real-time and intelligent processing. With the wide applications of convolutional neural networks in recent years, many deep learning based methods [1,5,6,10,21,28,30] are proposed, which can be trained in an end-to-end manner thanks to the public paired shadow removal datasets [28, 32]. Compared with traditional methods, deep learning methods can acquire better results with less inference time consuming. However, due to the continuous illumination change in nature, it is hard to acquire accurate paired shadow and shadow-free images. Therefore, existing datasets for shadow removal lack diversity and scale and impede the generalization ability of networks.

To overcome the limitation of existing shadow removal datasets, some studies attempt to train the network in an unsupervised manner [22, 32] with unpaired shadow and shadow-free images. Yet, large spectral gaps may exist between the domain of shadow and shadow-free images. Models trained on such datasets may cause spectral distorted results when applied to the real-world shadow images. Some studies [21, 23] then generate shadow and shadow-free patches from existing shadow images, which may need carefully filter out shadow patches with the guidance of shadow masks. Another solution is enlarging the paired shadow removal dataset with a shadow simulation model [18]. However, these works are based on the statistics of existing shadow removal datasets which cannot cover all kinds of shadow patterns.

Inspired by that single-band shadow simulation can get rid of complex relations among different bands [18], we propose a simple yet robust single-band shadow simulation method. Specifically, our proposed shadow simulation method is a simple linear model whose parameters are estimated for each given shadow-free image and shadow mask. In such manner, we can apply our shadow simulation method to large amounts of shadow-free images

*Corresponding Author

with arbitrary masks to generate corresponding shadow images and acquire a large-scale training dataset. Furthermore, we propose a two-stage framework, Gray-to-Color Shadow Removal Network (G2C-DeshadowNet) to decouple the shadow removal tasks into single-band (grayscale) shadow removal and grayscale image colorization. To make full use of the global information of an image, we adopt modified self-attention blocks [7] in our network design. For the first stage, given an shadow image, our Grayscale Enhancement Network takes as input a grayscale shadow image and aims to generate the grayscale shadow-free image with the gudiance of shadow mask. Then the Colorization Network attempts to colorize the output of Grayscale Enhancement Network with the guidance of residual information from colored shadow image. Our contributions are summarized as follows:

• We propose a simple yet robust shadow simulation method to generate a large-scale paired simulated shadow dataset from the Places2 dataset [39], with which, the trained model gains better generalization ability.

• We propose a two-stage framework to decopule the shadow remvoal task into grayscale shadow removal and shadow regions colorization.

• The proposed shadow simulation and removal framework acquires favorable shadow removal results against state-of-the-art methods quantitatively and qualitatively on ISTD+ [20], SRD [28] and SBU [31] datasets.

2. Related Work

2.1. Shadow Removal with Existing Shadow Datasets

Traditional methods remove shadows by formulating physical models with prior knowledge of shadows, such as image gradient [9,12] and illumination information [33,37]. In recent years, with the availability of paired shadow removal datasets [28, 32], many studies have applied deep learning methods [1, 5, 6, 10, 21, 28] to the shadow removal task by learning powerful features in an end-to-end manner, contributing to a significant improvement from traditional methods. Qu *et al*. [28] attempted to acquire shadow matte for shadow removal with a multi-context embedding network. Inspired by the stacked generative adversarial network [36], Wang *et al*. [32] introduced a jointly shadow detection and removal framework. Taking into consideration of non-shadow information, Chen *et al*. [1] proposed a contextual patch matching and transfer module to restore the shadow regions with shadow-free regions as reference. However, they heavily depended on the artificial design of the patch matching step and their results are easily suffered from blurriness. Le *et al*. [20] introduced a shadow image decomposition workflow for shadow removal. Fu *et al*. [10] further proposed an auto-exposure fusion model to deal with shadows on a single image. Despite that the above deep learning methods acquire satisfying results, most of existing methods [10,20,28,32] seldom utilize the non-local information, which can further facilitate image restoration.

In our framework, we acquire non-local information by inserting the patch self-attention module. Such network design is beneficial for shadow removal, as demonstrated in the ablation study. Besides, to reduce the solution space, we proposed a two-stage framework to decompose shadow removal into grayscale shadow removal and colorization.

2.2. Shadow Removal with Extensive Shadow Dataset

Existing shadow removal datasets have limited amounts of training data and lack diverse scene and shadow patterns. Models trained with these datasets will be short in generalization ability. To get rid of the limiation of existing paired datasets, unsupervised methods are proposed [16, 21–23]. Hu *et al*. [16] proposed to remove shadows together with generating shadows, which can be trained with unpaired shadow and shadow-free images. Liu *et al*. [22] further improved [16] by learning the light-related feature maps to guide the removal of shadow. However, large gaps may exist between the domains of shadow and shadow-free images. Shadow removal results of these models tend to have spectral distortion issue. Le *et al*. [21] cropped shadow and shadow-free patches from the same shadow images as the training data to reduce the domain gaps between shadow and shadow-free images. However, they use the shadow region as prior and suffer from a high computational load.

The other line of works is to enlarge the paired shadow removal datasets by simulating shadow images [5, 18, 23]. Inoue *et al*. [18] obtained the the prior information of shadows based on the statistics of ISTD dataset, and simulated shadows on the collected shadow-free images with the prior knowledge. The enlarged dataset can facilitate the model's performance. Cun *et al*. [5] trained a shadow simulation network with ISTD dataset to generate realistic shadow images from shadow-free images given the shadow mask. Liu *et al*. [23] proposed a CycleGAN-based method to generate shadows on the shadow-free regions of a shadow image under the guidance of a mask. Although the scale and diversity of the dataset are enlarged, the generated shadows have similar patterns to those in the ISTD dataset [18]. Models trained with such enlarged datasets may not deal with shadow patterns outside ISTD dataset.

Towards the real-world shadow removal with unknown shadow patterns, we introduce a robust yet straightforward shadow simulation strategy to make shadow simulation on the grayscale to reduce the complex relation modeling among different channels. With our shadow simulation strategy, we can generate a large-scale and diverse shadow removal training dataset. This shadow simulation dataset

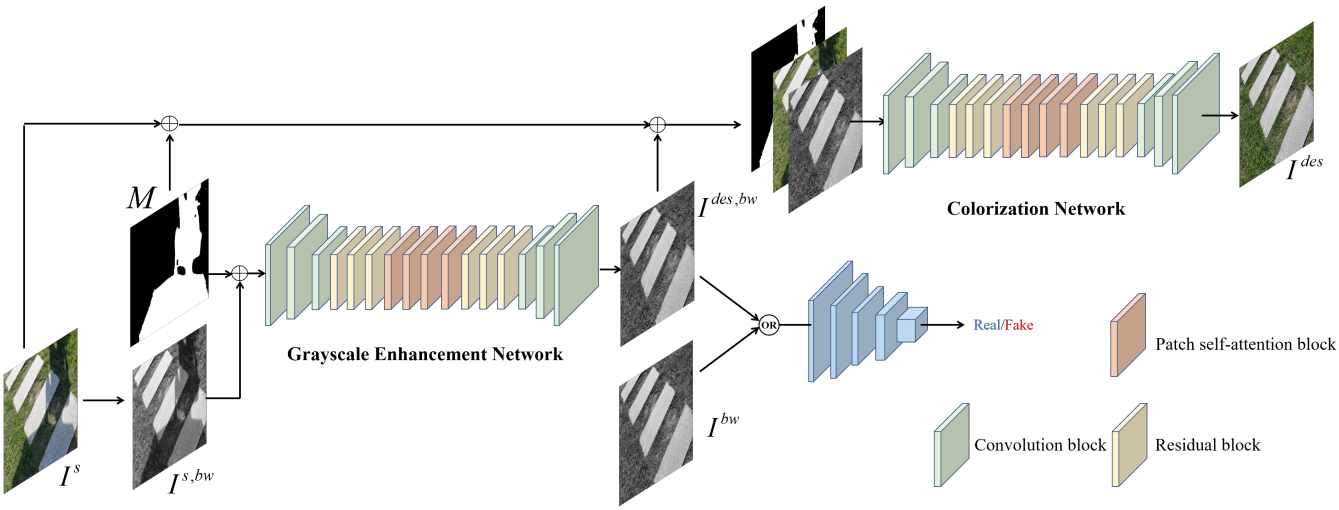


Figure 1. Overview of our two-stage framework. At the first stage, an RGB image is first transformed into grayscale and then concatenated with the shadow mask as the input of the Grayscale Enhancement Network to generate the grayscale shadow-free image. At the second stage, shadow mask, shadow-free grayscale image, and color shadow image serve as the Colorization Network's input to colorize the estimated grayscale shadow-free into RGB shadow-free image.

empowers our proposed two-stage framework adapting well to the complex real-world scenario.

3. Method

Shadows have various patterns on different bands of an image, and the relationships among different bands are hard to establish [18]. Considering the above nature of shadows, our proposed framework decomposes the shadow removal task on RGB images into two stages: grayscale shadow removal and colorization. At the first stage, we first transform a color shadow image into grayscale and then restore it into a grayscale shadow-free image by a Grayscale Enhancement Network. At the second stage, we design a Colorization Network to colorize the grayscale shadow-free image with the guidance of residual color information. In such manner, our proposed method avoids modeling shadow relations among different bands. The framework is illustrated in Figure 1. Besides, we introduce a shadow simulation method to generate a large-scale training dataset to adapt for real-world shadow removal.

3.1. Grayscale Enhancement Network

At this stage, we restore the shadow image into the grayscale shadow-free image by the Grayscale Enhancement Network, which contains an encoder, several patch self-attention blocks, and a decoder. The encoder first encodes the input grayscale image and shadow mask to high-level features. The patch self-attention blocks then take the high-level features to acquire contextual features empowered with global information. Finally, the contextual features are sent to the decoder to reconstruct the image.

The encoder consists of three groups of Conv-BatchNormalization-ReLU and three residual blocks [15]. The decoder part has three residual blocks followed by three groups of Deconv-BatchNormalization-ReLU.

For the patch self-attention blocks, we draw the self-attention idea from [35] and modify their self-attention module into patch self-attention module to fit the shadow removal task. Patch self-attation blocks can reduce the computation cost and are suitable for the restoration task. The feature maps from the encoder are first cropped into patches and flattened into 1-D arrays. Three convolution layers then process these 1-D arrays to acquire Key matrix, Query matrix, and Value matrix for self-attention operation and finally reshaped back to the original shape. The patch self-attention module is illustrated in Figure 2. The patch self-attention block is stacked by four times to fully make use of non-local information.

Given an RGB image, we first transform it into the grayscale image by weighted summation across the channels:

$$I^{s,bw} = 0.31 * I^{s}(R) + 0.59 * I^{s}(G) + 0.10 * I^{s}(B), \quad (1)$$

where I^{s} and $I^{s,bw}$ denote the color shadow image and grayscale shadow image, respectively. R, G and B are three channels of I^{s}. Then we formulate the shadow removal on grayscale image as:

$$I^{des,bw} = G_1(I^{s,bw}, M), \quad (2)$$

where G_1 denotes the Grayscale Enhancement Network whose output is the grayscale shadow-free image. M is the shadow mask.

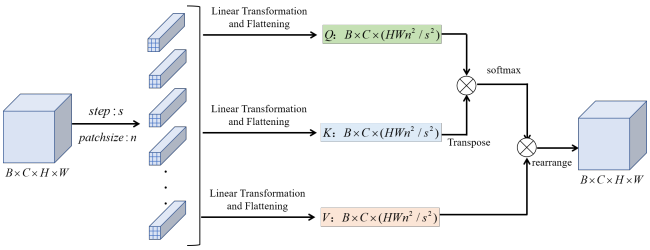


Figure 2. Patch self-attention block.

Two loss functions are adopted in this stage. The first is the pixel-level loss function, L1 loss :

$$L_{pixel}^{bw} = \mathrm{E}[||I^{des,bw} - I^{bw}||_1], \tag{3}$$

where I^{bw} denotes the grayscale ground truth.

The second is the generative adversarial loss function which aims to reconstruct the detailed information:

$$L_{adv}^{bw} = \mathrm{E}[log(1 - D_1(I^{des,bw}))] + \mathrm{E}[log(D_1(I^{bw}))]. \tag{4}$$

For the designation of D_1, we adopt the discriminator from Patch-GAN [19] and replace the batch-normalization with spectral-normalization [24].

The total loss function can be described as:

$$L_{total}^{bw} = L_{adv}^{bw} + \omega_1 L_{pixel}^{bw}, \tag{5}$$

where ω_1 is a hyperparameter. We empirically set $\omega_1 = 100$.

3.2. Colorization Network

Colorization Network aims to colorize the grayscale shadow-free image obtained from the first stage, with the guidance of residual information from the color shadow image. Colorization Network adopts the same network structure with Grayscale Enhancement Network except the input layer. The colorization process can be formulated as:

$$I^{des} = G_2(I^s, I^{des,bw}, M), \tag{6}$$

where G_2 denotes the Colorization Network which aims to acquire color shadow-free result I^{des}.

For the model trained on our simulation dataset, to avoid the negative effect of simulation input, the colorization process is modified as :

$$I^{des} = G_2((1 - M) \cdot I^s, I^{des,bw}, M). \tag{7}$$

The simulation process will be introduced in the Sec. 3.3.

We use the loss function L_{total}^{rgb} to optimize our Colorization Network. :

$$L_{total}^{rgb} = \mathrm{E}[||I^{des} - I||_1]. \tag{8}$$

3.3. Shadow Simulation

As mentioned in the Sec.1, existing shadow removal datasets lack scale and diversity, constraining the generalization ability of models trained on them. We introduce a simple yet robust shadow simulation method to generate large amounts of shadow images from shadow-free images and arbitrary masks for facilitating our G2C-DeshadowNet training.

According to [29], pixels in an image can be determined by the illumination and reflectance of corresponding position. Pixels in non-shadow region can be expressed as:

$$I_x^{ns}(\lambda) = L_x^d(\lambda)R_x(\lambda) + L_x^a(\lambda)R_x(\lambda). \tag{9}$$

While pixels in shadow region can be expressed as:

$$I_x^s(\lambda) = \alpha_x L_x^a(\lambda)R_x(\lambda). \tag{10}$$

$L_x^d(\lambda)$, $L_x^a(\lambda)$, α_x and $R_x(\lambda)$ are the direct illumination, ambient illumination, shadow matting and reflectance of pixel x on λ band, respectively. The relationship between the same pixel under shadow and non-shadow condition can be described as:

$$I_x^{ns}(\lambda) = \frac{1}{\alpha_x} I_x^s(\lambda) + L_x^d(\lambda)R_x(\lambda), \tag{11}$$

which is a classic linear model.

Given the shadow mask M, Equation 11 can be further described as [37]:

$$\frac{I_x^s(\lambda) - \mu(M \cdot I^s(\lambda))}{\sigma(M \cdot I^s(\lambda))} = \frac{I_x^{ns}(\lambda) - \mu(M \cdot I^{ns}(\lambda))}{\sigma(M \cdot I^{ns}(\lambda))}. \tag{12}$$

Let denote $k(\lambda) = \frac{\mu(M \cdot I^s(\lambda))}{\mu(M \cdot I^{ns}(\lambda))}$, then Equation 12 can be simplified as:

$$I_x^s(\lambda) = k(\lambda)^2 I_x^{ns}(\lambda) + (k(\lambda) - k(\lambda)^2)\mu(M \cdot I^{ns}(\lambda)). \tag{13}$$

Given the shadow mask M, we only need $k(\lambda)$ to simulate shadows on the band λ. Hence, three parameters are required to simulate shadows on RGB images. However, complex relationships among shadows in three bands exist according to [18], and it is hard to model the relations between the parameters. To simplify the simulation, we make the shadow simulation on the grayscale, which only needs one parameter:

$$I_x^{s,bw} = k^2 I_x^{ns,bw} + (k - k^2)\mu(M \cdot I^{ns,bw}). \tag{14}$$

During training, k is empirically randomly sampling between 0.2 and 0.8.

(a) (b)

Figure 3. An example of simulated shadow images (a) and adjusted input (b). From left to right are the input, indicating mask and simulated result/adjust input, respectively.

It's worth noting that there still exists domain gaps between simulated grayscale shadow images and real grayscale shadow images. To reduce the gap, we further adjust shadow regions of real grayscale shadow images and generate the adjusted input. We assume that shadow regions share a similar distribution with shadow-free surrounding regions. We extract the shadow-free surrounding regions as follows. The original shadow mask M is first empirically dilated by 7 pixels as M^{d7}. The different set between M^{d7} and M is denoted as the surrounding shadow-free regions, M^d. The pixel values of shadow regions is adjusted under the guidance of M^d:

$$I_x^{s,bw1} = \sqrt{\frac{\mu(M^d \cdot I^{s,bw})}{\mu(M \cdot I^{s,bw})}}(I_x^{s,bw} - \mu(M \cdot I^{s,bw})) + \mu(M^d \cdot I^{s,bw}). \tag{15}$$

$$I_x^{s,bw2} = k_0^2 I_x^{s,bw1} + (k_0 - k_0^2)\mu(M \cdot I^{s,bw1}). \tag{16}$$

During inference, k is set as a fixed parameter $k_0 = 0.64$.

We simulate all shadow images by shadow-free images from the Places2 dataset [39], which is rich in scale and diversity. With our simulation strategy, our model can be trained without real-world paired shadow dataset and applied for real-world shadow removal. An example of simulated image and adjusted input are shown in Figure 3 (a) and (b), respectively.

4. Experiment

4.1. Experiment Settings

We conduct extensive experiments on several benchmark datasets, including ISTD+ dataset [20,32], SRD dataset [28] and SBU dataset [31], to compare our G2C-DeshadowNet with state-of-the-art methods. For the SRD dataset, We extract shadow masks from the differential map between shadow and shadow-free images by Otsu [27] threshold, which will be used for training and testing. SBU is a shadow detection dataset and we use it for generalization ability evaluation on real-world images.

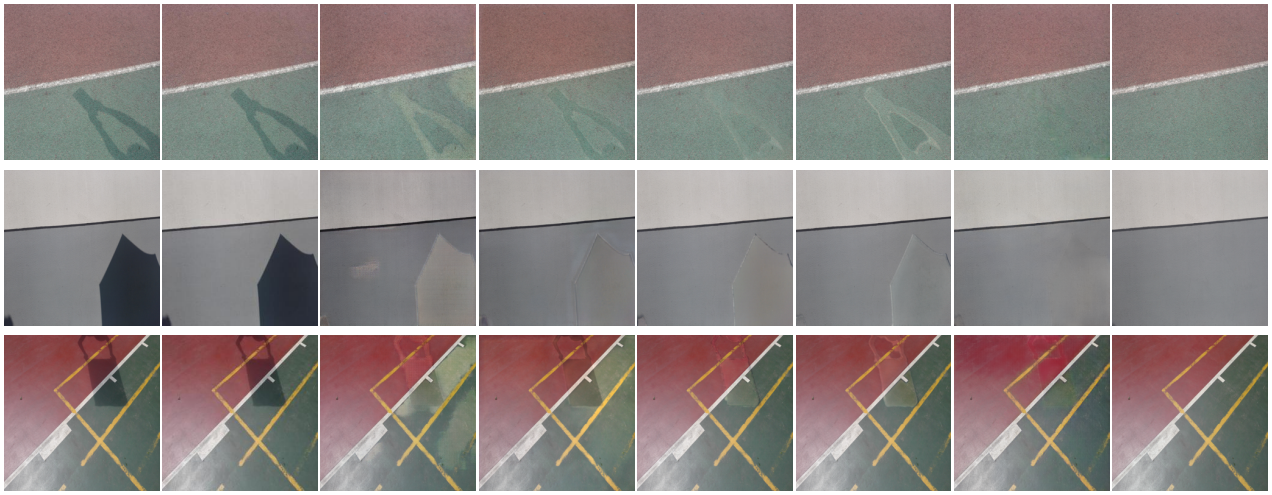
Evaluation Method. We resize the results to 256*256 and compute the RMSE in the LAB color space like [20, 23].

Implementation details. Adam optimizer is adopted to optimize our model's parameters with a batch size of 1. First momentum, second momentum and weight decay of Adam optimizer are set as 0.9, 0.99 and 0.001, respectively. The learning rate is set as 2×10^{-4}. The training epoch is set as 200.

4.2. Comparison with Models Trained on ISTD+ Dataset

For fair comparisons, this section only compares supervised methods trained on the ISTD+ dataset [20], including Yang [34], Gong [11], Guo [14], ST-CGAN [32], DSC [17], SP+M-Net [20], CANet [1] and AEF [10]. The first first methods need no shadow mask while the last four methods use shadow masks as reference in the shadow removal process.All the compared results in the paper are obtained from the corresponding authors, generated by their provided models or from their original papers. Table 1 shows the quantitative results. Due to the lack of shadow masks, methods without guidance of shadow masks show inferior results compared with methods with guidance of shadow masks. In the methods with shadow masks guidance, our method gains the best RMSE scores in shadow regions and the whole image and the second best RMSE scores in non-shadow regions. Specifically, our method outperforms the best two models, AEF [10] and SP+M-Net [20], by 1.5% and 19.0%, respectively, in shadow areas, and surpasses other methods by an even more significant extent. Note that *Gong [11] is an interactive method which only modifies the information of shadow regions and retains all shadow-free information from shadow images.. *Vasluianu [30] does not publish their codes. We reproduce and improve their model by concatenating the shadow masks to input.

Figure 4 shows some visual results. Our method can completely remove shadows and has better consistent brightness and smoothness between shadow and non-shadow regions. While the traditional method [14] fails to remove shadows completely with several remaining artifacts. The possible reason is that it lacks global informa-



(a) Input (b) Guo [14] (c) ST-CGAN [32] (d) DSC [17] (e) SP+M-Net [20] (f) G2R [23] (g) Ours (h) GT

Figure 4. Visual results on ISTD+ dataset. All models are trained in a supervised manner on ISTD+ dataset.

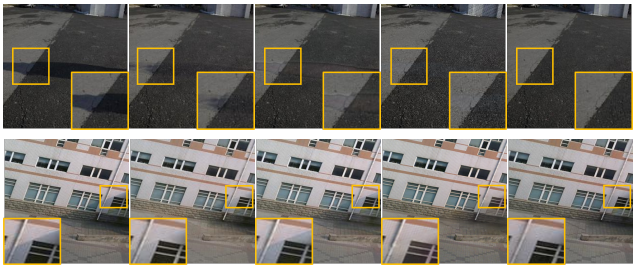
Method	Mask	Shadow	Non-shadow	All
Yang [34]	No	24.7	14.3	15.9
Guo [14]	No	22.0	3.1	6.1
ST-CGAN [32]	No	13.4	7.9	8.6
DSC [17]	No	9.2	6.3	6.6
*Gong [11]	No	13.3	**2.6**	4.3
SP+M-Net [20]	Yes	7.9	3.1	3.9
*Vasluianu [30]	Yes	7.4	3.1	3.7
CANet [2]	Yes	8.9	6.1	6.2
AEF [10]	Yes	6.5	3.8	4.2
Ours	Yes	**6.4**	2.9	**3.5**

Table 1. Quantitative results on ISTD+ dataset. The evaluation metric is RMSE. The lower score indicates the better results.

Method	Shadow	Non-shadow	All
DeshadowNet [28]	11.8	4.8	6.6
DHAN [5]	8.9	4.7	5.7
DSC [17]	10.9	5.0	6.2
AEF [10]	8.6	5.8	6.5
Ours	**8.4**	**4.6**	**5.5**

Table 2. Quantitative results on SRD dataset. Lower score means better performance. The best results are highlighted in bold.

tion for scene understanding. Deep learning based methods, *e.g*. SP+M-Net [20], can acquire better results compared with traditional methods. However, their results usually suffer from inconsistency between shadow regions and shadow-free regions. Results of ST-CGAN [32] suffer from severe color distortion and blurriness. The results of DSC [17], exists obvious shadow boundaries between shadow and shadow-free regions.



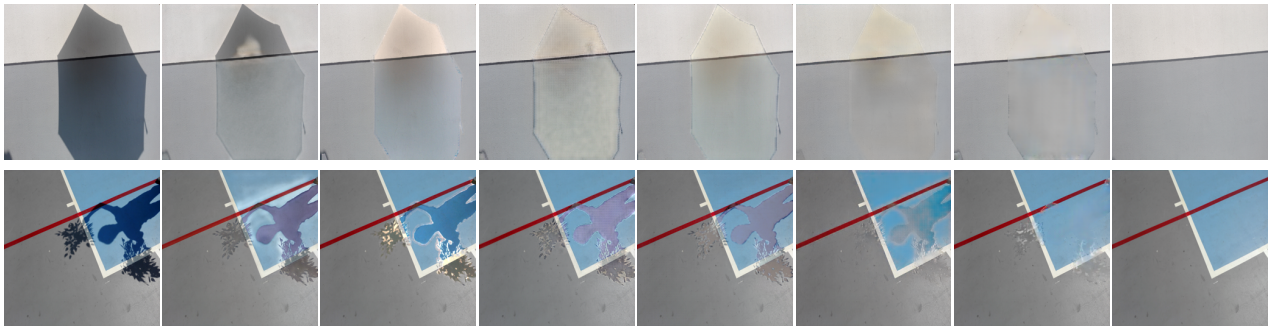
(a) Input (b) DHAN [5] (c) DSC [17] (d) Ours (e) GT

Figure 5. Shadow Results on SRD dataset.

4.3. Comparison with Models Trained on SRD Dataset

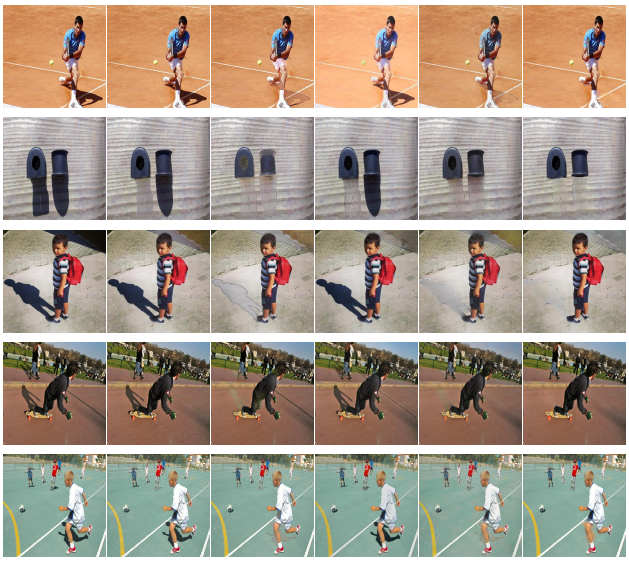
For fair comparisons on SRD dataset [28], we compare our model with models trained on SRD dataset, including DeshadowNet [28], AEF [10], DSC [17] and DHAN [5].

Table 2 reports the quantitative results on the SRD dataset. The proposed method gains the best results in shadow regions, shadow-free regions and all regions compared with existing methods. Specifically, compared with corresponding best methods (i.e., DSC, and DHAN), our method reduces the RMSE scores of shadow region, shadow-free region and all regions by 2.3%, 2.1% and 3.5%, respecitvely. Figure 5 shows visual results on the SRD dataset. DSC [17] and DHAN [5] remain shadow artifacts in their results. In contrast, our method nearly perfectly removes all shadows.



(a) Input (b) LGSN [22] (c) Le *et al*. [21] (d) G2R [23] (e) SP+M-Net [20] (f) Ours* (g) Ours (h) GT

Figure 6. Shadow results on ISTD+ dataset. The comparison models are either trained in an unsupervised/weekly-supervised manner on ISTD+ dataset. Ours* means our model trained on ISTD+ dataset. Ours is the model trained on simulated shadow dataset generated from shadow-free images of Places2 dataset [39].



(a) Input (b) (c) (d) (e) (f)

Figure 7. Visual results beyond ISTD+ dataset. The compared models are trained on ISTD+ dataset. From (b) to (f) are the results of SP+M-Net [20], DHAN [5], G2R [23], our method trained on ISTD+ dataset in a supervised manner, and our method trained on the simulated shadow dataset.

Method	Dataset	Shadow	Non-shadow	All
Ours*	Paired	6.4	2.9	3.5
MaskShadowGAN [16]	Unpaired	10.8	3.8	4.8
LGSN [22]	Unpaired	9.8	3.4	4.4
Le *et al*. [21]	-	9.7	**3.0**	4.0
G2R† [22]	-	8.8	3.6	4.5
DHAN+DA *et al*. [5]	Paired	11.2	7.0	7.8
Ours	-	**8.6**	3.2	4.2

Table 3. Quantitative results on ISTD+ dataset. The other methods are trained with unsupervised/weakly-supervised manner on ISTD+ dataset or with extensive datasets. Ours* denotes that our model trained on ISTD+ dataset. †Note we adopt the original output of their model [22] for fair comparison just like other methods.

4.4. Effectiveness of Shadow Simulation Strategy

This section quantitatively and qualitatively demonstrates the effectiveness of our shadow simulation strategy. We first compare our model trained on our simulated shadow dataset and the ISTD+ dataset as well as several unsupervised or weakly supervised methods or methods trained on extensive dataset, including G2R-ShadowNet [23], Le *et al*. [21], DHAN+DA [5] and LGSN [22]. The evaluation is performed on the ISTD+ dataset. As shown in Table 3, our model trained on the shadow simulation dataset obtains comparable results with that trained on the ISTD+ dataset and outperforms all other methods in shadow regions by a large margin. It's worth noting that our model trained on our shadow simulation dataset doesn't use any images from ISTD+ dataset, while all the compared methods use at least the shadow images or shadow-free images from the ISTD+ dataset. Figure 6 displays some visual results. Our model trained on the simulation dataset can well remove the shadows and restore the images more naturally, even outperforming methods trained on the ISTD+ dataset. It indicates that our shadow simulation dataset is beneficial for shadow removal.

To further validate the generalization ability of our model trained with the proposed simulation shadow dataset, we compare our model trained on the simuated shadow dataset with those trained on ISTD+ dataset and test on the real-world SBU dataset [31]. We also compare our models with G2R-ShadowNet [23], DHAN [5] and SP+M-Net [20], which are trained on ISTD+ dataset. The visual results are displayed in Figure 7. We observe that G2R-ShadowNet [23], SP+M-Net [20] can hardly remove shadows in the im-

ages outside ISTD+ dataset. DHAN [5] can partially remove the shadows and remain some residual shadow artifacts. Our method trained on ISTD+ dataset has less artifacts than other methods. Our model trained by simulated shadow dataset can perfectly remove shadows in unknown cases, indicating that our model design and shadow simulation strategy can contribute to better generalization ability.

4.5. Ablation Study

We first conduct ablation studies on ISTD+ dataset to validate the contribution of network design.

• G2C-DeshadowNet w/o D_1 : removing the discriminator D_1.

• G2C-DeshadowNet w/o G_1 : removing Grayscale Enhancement Network G_1 and discriminator D_1, to evaluate the effectiveness of our two-stage design.

• G2C-DeshadowNet w/o PST : removing all patch self-attention blocks.

• G2C-DeshadowNet w/ $1PST$: retaining only one patch self-attention block in both Grayscale Enhancement Network and Colorization Network.

The results are reported in Table 4. G2C-DeshadowNet w/o G_1 performs poorly in both restoration of shadow region and retainment of shadow-free region, confirming that our two-stage design can greatly improves the shadow removal results. G2C-DeshadowNet w/o D_1 works a little worse than the full model, indicating that D_1 contributes positively to our model. The accuracy of G2C-DeshadowNet w/o PST is lower than G2C-DeshadowNet w/ 1PST, which means that the patch self attention block can make full use of non-local information to restore the shadow region. G2C-DeshadowNet w/ 1PST has lower scores than the full model, which confirms that stacking patch self-attention blocks improves the shadow removal results.

We then conduct ablation studies for our shadow simulation method. The variant models are trained on simulated shadow dataset and test on the ISTD+ dataset.

• G2C-DeshadowNet w/o CA : removing color adjustment introduced in Equation 16, and directly process the real shadow images.

The evaluation results are reported in Table 5. Our full shadow simulation and removal model outperforms G2C-DeshadowNet* w/o CA by a large extent. The reason is that shadows simulated by Equation 14 still have domain gaps with real-world shadows. With our color adjustment step to alleviate the domain gap issue, the adjusted real-world shadow images can adapt to our model.

Method	Shadow	Non-shadow	All
G2C-DeshadowNet w/o D_1	6.7	3.2	3.8
G2C-DeshadowNet w/o G_1	8.4	5.5	6.0
G2C-DeshadowNet w/o PST	7.5	3.1	3.9
G2C-DeshadowNet w/ 1PST	6.6	3.2	3.8
Ours*	**6.4**	**2.9**	**3.5**

Table 4. Ablation studies of each component in our model. The models are trained and tested on ISTD+ dataset.

Method	Shadow	Non-shadow	All
G2C-DeshadowNet w/o CA	16.0	3.3	5.5
Ours	**8.6**	**3.2**	**4.2**

Table 5. Ablation studies for our shadow simulation strategy. The models are trained on simulated shadow dataset and tested on ISTD+ dataset.

(a) (b) (c) (d)

Figure 8. Failure Cases. The first row shows some real-world shadow images and the second row shows their results generated by our model trained on the simulated shadow dataset. Our method may fail in very dark shadow removal.

5. Conclusion

In this paper, we propose a robust shadow removal model, *i.e.*, G2C-DeshadowNet. The proposed model first removes shadows in the grayscale and then colorizes the grayscale shadow-free images with the guidance of color shadow images. This workflow design improves the accuracy and robustness of the model. Stacked patch self-attention blocks are introduced in our model to further utilize non-local information for more accurate restoration results. The proposed method obtains state-of-the-art results on ISTD+, SRD and SBU dataset. Moreover, we introduce a novel shadow simulation strategy, with which our framework can be trained without real-world shadow removal dataset and applied to real-world shadow removal. The shadow simulation strategy can further improve the generalization ability of our model. Although our method work well in most cases, it may fail when the shadow regions are too dark as shown in Figure 8. In such case, our model may be hard to acquire effective information to restore the image. In future work, we will attempt to establish more efficient shadow simulation model and more powerful network for very dark shadow removal.

References

[1] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Canet: A context-aware network for shadow removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4743–4752, 2021. 1, 2, 5

[2] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Canet: A context-aware network for shadow removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4743–4752, October 2021. 6

[3] Yung-Yu Chuang, Dan B Goldman, Brian Curless, David H Salesin, and Richard Szeliski. Shadow matting and compositing. In *ACM SIGGRAPH 2003 Papers*, pages 494–500. 2003. 1

[4] Rita Cucchiara, Costantino Grana, Massimo Piccardi, Andrea Prati, and Stefano Sirotti. Improving shadow suppression in moving object detection with hsv color information. In *ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No. 01TH8585)*, pages 334–339. IEEE, 2001. 1

[5] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10680–10687, 2020. 1, 2, 6, 7, 8

[6] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10213–10222, 2019. 1, 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[8] Graham D Finlayson, Mark S Drew, and Cheng Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85(1):35–57, 2009. 1

[9] Graham D Finlayson, Steven D Hordley, Cheng Lu, and Mark S Drew. On the removal of shadows from images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):59–68, 2005. 1, 2

[10] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10571–10580, 2021. 1, 2, 5, 6

[11] Han Gong and Darren Cosker. Interactive removal and ground truth for difficult shadow scenes. *JOSA A*, 33(9):1798–1811, 2016. 5, 6

[12] Maciej Gryka, Michael Terry, and Gabriel J Brostow. Learning to remove soft shadows. *ACM Transactions on Graphics (TOG)*, 34(5):1–15, 2015. 2

[13] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Single-image shadow detection and removal using paired regions. In *CVPR 2011*, pages 2033–2040. IEEE, 2011. 1

[14] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2956–2967, 2012. 1, 5, 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[16] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2472–2481, 2019. 2, 7

[17] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7454–7462, 2018. 5, 6

[18] Naoto Inoue and Toshihiko Yamasaki. Learning from synthetic shadows for shadow detection and removal. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. 1, 2, 3, 4

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 4

[20] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8578–8587, 2019. 2, 5, 6, 7

[21] Hieu Le and Dimitris Samaras. From shadow segmentation to shadow removal. In *European Conference on Computer Vision*, pages 264–281. Springer, 2020. 1, 2, 7

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

[23] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4927–4936, 2021. 1, 2, 5, 6, 7

[24] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 4

[25] Ankit Mohan, Jack Tumblin, and Prasun Choudhury. Editing soft shadows in a digital photograph. *IEEE Computer Graphics and Applications*, 27(2):23–31, 2007. 1

[26] Sohail Nadimi and Bir Bhanu. Physical models for moving shadow and object detection in video. *IEEE transactions on pattern analysis and machine intelligence*, 26(8):1079–1087, 2004. 1

[27] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979. 5

[28] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4067–4075, 2017. 1, 2, 5, 6

[29] Yael Shor and Dani Lischinski. The shadow meets the mask: Pyramid-based shadow removal. In *Computer Graphics Forum*, volume 27, pages 577–586. Wiley Online Library, 2008. 4

[30] Florin-Alexandru Vasluianu, Andrés Romero, Luc Van Gool, and Radu Timofte. Shadow removal with paired and unpaired learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 826–835, 2021. 1, 5, 6

[31] Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *European Conference on Computer Vision*, pages 816–832. Springer, 2016. 2, 5, 7

[32] Jifeng Wang, Xiang Li, and Jian Yang. 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*, pages 1788–1797, 2018. 1, 2, 5, 6

[33] Chunxia Xiao, Donglin Xiao, Ling Zhang, and Lin Chen. Efficient shadow removal using subregion matching illumination transfer. In *Computer Graphics Forum*, volume 32, pages 421–430. Wiley Online Library, 2013. 1, 2

[34] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. *IEEE Transactions on Image processing*, 21(10):4361–4368, 2012. 5, 6

[35] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 3

[36] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[37] Ling Zhang, Qing Zhang, and Chunxia Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing*, 24(11):4623–4636, 2015. 1, 2, 4

[38] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. Distraction-aware shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2019. 1

[39] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 2, 5, 7