Efficient Model-Driven Network for Shadow Removal

Yurui Zhu[1], Zeyu Xiao[1], Yanchi Fang[2], Xueyang Fu[1] *, Zhiwei Xiong[1], Zheng-Jun Zha[1]

[1] University of Science and Technology of China, China
[2] University of Toronto, Canada
{zyr, zeyuxiao}@mail.ustc.edu.cn, yanchi.fang@mail.utoronto.ca, {xyfu, zwxiong, zhazj}@ustc.edu.cn

Abstract

Deep Convolutional Neural Networks (CNNs) based methods have achieved significant breakthroughs in the task of single image shadow removal. However, the performance of these methods remains limited for several reasons. First, the existing shadow illumination model ignores the spatially variant property of the shadow images, hindering their further performance. Second, most deep CNNs based methods directly estimate the shadow free results from the input shadow images like a black box, thus losing the desired interpretability. To address these issues, we first propose a new shadow illumination model for the shadow removal task. This new shadow illumination model ensures the identity mapping among unshaded regions, and adaptively performs fine grained spatial mapping between shadow regions and their references. Then, based on the shadow illumination model, we reformulate the shadow removal task as a variational optimization problem. To effectively solve the variational problem, we design an iterative algorithm and unfold it into a deep network, naturally increasing the interpretability of the deep model. Experiments show that our method could achieve SOTA performance with less than half parameters, one-fifth of floating-point of operations (FLOPs), and over seventeen times faster than SOTA method (DHAN).

Introduction

Shadows, which are caused by light being blocked by objects, widely exist in various natural scenes. Shadows present a substantial challenge for computer vision applications such as tracking (Sanin, Sanderson, and Lovell 2010; Guo et al. 2020) and object detection (Cucchiara et al. 2003). Consequently, removing shadows is a crucial preprocessing step in computer vision, and has attracted extensive research attention. Currently, deep CNNs-based methods (Qu et al. 2017; Wang, Li, and Yang 2018; Cun, Pun, and Shi 2020; Le and Samaras 2020; Liu et al. 2021c) dominate this field and achieve state-of-the-art performance. However, these approaches have some inherent problems:

Simplified assumption in the existing shadow illumination model. Recently, many researchers try to further explore the physical properties of the shadow phenomenon. In (Le and Samaras 2019, 2020), the illumination model of the

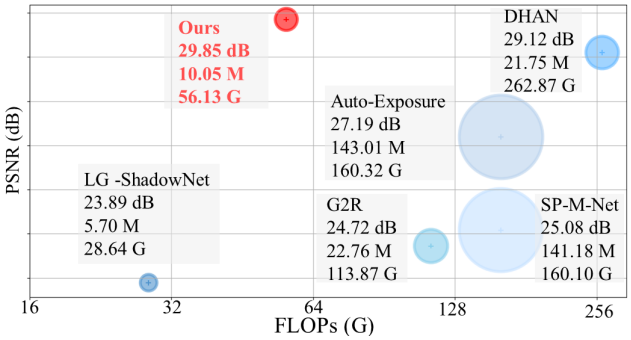

Figure 1: Comparison among different models PSNR performance, number of models parameters and FLOPs. Shadow removal results are achieved on ISTD dataset.

transformation relationship between the shadow-free image $\mathbf{I}_{ns}$ and shadow image $\mathbf{I}_s$ is explicitly expressed as

$$\mathbf{I}_{lit} = \omega * \mathbf{I}_s + b, \tag{1}$$

$$\mathbf{I}_{ns} = \boldsymbol{\alpha} * \mathbf{I}_s + (\mathbf{1} - \boldsymbol{\alpha}) * \mathbf{I}_{lit}, \tag{2}$$

where $\mathbf{I}_{lit}$ is the relit result of transforming $\mathbf{I}_s$ with six constants ($\omega = [\omega_R, \omega_G, \omega_B], b = [b_R, b_G, b_B]$) for different image channels. The matte $\boldsymbol{\alpha}$ balances the shadow effects in penumbra areas. Given the above formula, they design networks to predict different parameters in two stages. In the first stage, this illumination model strongly assumes that all areas of the umbra are equally affected by shadows, which largely ignores the spatial-variant property of shadow images. We argue that directly predicting a uniform mapping function for all umbra area pixels is not accurate. Neither of these methods considers a more finer-grained, more adaptive mapping function for each pixel in umbra regions.

Lacking sufficient interpretability of CNNs. Deep CNNs significantly promote the single image shadow removal performance in conjunction with the large-scale datasets of shadow/non-shadow pairs. However, many deep CNNs-based methods (Qu et al. 2017; Wang, Li, and Yang 2018; Cun, Pun, and Shi 2020) mainly focus on designing the network architectures, and then adopt an end-to-end training strategy to directly learn the mapping function between $\mathbf{I}_s$ and $\mathbf{I}_{ns}$ to remove shadows. This learning paradigm makes the deep network act as a black box and ignores the intrin-

*Corresponding author.

sic prior knowledge of the shadow image itself, resulting in weak interpretability and limited performance.

Insufficient use of shadow mask information. Current high-quality datasets for the shadow removal task, *e.g.*, Image Shadow Triplets Dataset (ISTD) (Wang, Li, and Yang 2018), have provided corresponding shadow masks. In order to utilize the shadow mask location information, many deep CNNs-based shadow removal methods (Qu et al. 2017; Wang, Li, and Yang 2018; Le and Samaras 2019, 2020; Liu et al. 2021c) directly concatenate the shadow image and mask as inputs, and then send them into the deep network. However, they may still mistake the dark albedo material areas like shadows, thus resulting in undesired artifacts in the results (as shown in Figure 5). Therefore, we argue that the mask information used to assist the network in locating shadow regions has not been fully exploited.

To alleviate the aforementioned issues, we propose an interpretable and effective deep network by combining the advantages of both model-driven methods and data-driven CNNs-based approaches. Specifically, we first propose a new shadow illumination model for this specific shadow removal task. Our proposed model integrates the shadow mask to design a constraint item, which could achieve the identity mapping among non-shadow regions pixels and adaptively fine-grained mapping between shadow region pixels and shadow-free counterparts within one stage. Secondly, based on the new illumination model, we reformulate the deshadow task as a variational optimization model, composed of the favorable data fidelity term and the shadow-related prior term. The favorable data fidelity ensures that estimated results are consistent with the proposed shadow illumination model, while the other terms learn shadow-relevant priors for removing shadow. To effectively solve the variational model, we design an optimization algorithm with the gradient descent strategy (Ruder 2016) and unfold it into deep CNNs. In this way, the operation process of our network is highly consistent with the optimization algorithm, thus the interpretability of the network has been well improved. Additionally, we construct a basic Dynamic Mapping Residual Block (DMRB), which conforms to the new shadow illumination model, to further improve the performance. The contributions of our paper are as follows:

- We propose a new illumination model at a fine-grained level for shadow removal. We verify the previous shadow illumination model neglects the spatially-variant property of the shadow images through statistical analysis. Our model is more comprehensive than the previous model, making the subsequent solution process more accurate.
- We propose an efficient model-driven network for shadow removal. Our network is built on the designed iterative optimization algorithm of shadow variational model, which largely increases its interpretability.
- We further design a Dynamic Mapping Residual Block (DMRB) as the basic module in our network, which is our proposed shadow illumination model-inspired. Compared with standard form Residual Block, DMRB can enhance our model performance without increasing additional parameters.
- Extensive experiments indicate that our method achieves leading shadow removal performance in terms of quantitative metrics, inference efficiency, and visual quality.

Related Work

Shadow Removal

Shadow removal is one of the fundamental tasks in the computer vision field. Traditional shadow removal methods (Shor and Lischinski 2008; Finlayson, Hordley, and Drew 2002; Wu, Zhang, and Kumar 2012) rely on the image intrinsic priors (*e.g.*, image gradients, illumination) and user interaction. For example, (Gong and Cosker 2014) designs the on-the-fly learning method by providing two rough user inputs for the pixels of the shadow and the lit area. Guo, Dai, and Hoiem (2012) combine pairwise relationships between shadow regions and non-shadow regions pixels to obtain illumination conditions.

Recently, there are many works to employ deep CNNs to remove shadows and achieve great breakthroughs. (Wang, Li, and Yang 2018; Qu et al. 2017) have proposed the public large-scale dataset (*e.g.*, ISTD) to train the network in an end-to-end manner. Ding et al. introduces the recurrent adversarial network to predict shadow mask and shadow-free images. (Hu et al. 2019a) novelly leverages direction-aware spatial context to promote shadow detection and removal performance. (Le and Samaras 2019) constructs a shadow illumination model and designs networks to remove shadow effects. (Cun, Pun, and Shi 2020) synthesizes realistic shadow images through adversarial learning to boost network performance. (Fu et al. 2021a) transfers shadow removal task as multi-exposure images fusion problem. Moreover, (Hu et al. 2019b; Le and Samaras 2020; Liu et al. 2021c) attempt to train their network with unpaired data.

Deep Unfolding Methods

Deep unfolding methods usually integrate the advantages of model-driven methods (*e.g.*, clearly interpretable) and the advantages of CNNs-based methods. (*e.g.*, powerful learning capability) They typically contain these steps: (1) unfold the specific iteration optimization algorithms, *e.g.*, iterative shrinkage-threshold algorithm (Beck and Teboulle 2009; Fu et al. 2019) and half-quadratic splitting algorithm (Afonso, Bioucas-Dias, and Figueiredo 2010; Zhang, Gool, and Timofte 2020); (2) parameterize the unrolled models; (3) update the learnable parameters via CNNs. Current many methods (Zhang et al. 2017; Dong et al. 2018; Mu et al. 2018; Li et al. 2019; Liu et al. 2019; Wang et al. 2020; Fu et al. 2021b; Liu et al. 2021a) in other computer vision tasks build different iterative optimization algorithms based on their physical models. Such as, dehazing method (Liu et al. 2019) unfolds the iterative algorithm, which solves the components of the well-known haze image physical formulation (Narasimhan and Nayar 2000; Fattal 2008), into CNNs.

However, such deep unfolding methods have not been explored on the shadow removal task. To our best knowledge, this is the first attempt to combine the model-driven optimization strategy with the CNNs-based method to achieve shadow removal.

Methodology

In this section, we illustrate our proposed shadow illumination model, the iterative solution to the minimization variational model, and the corresponding unfolding networks.

Shadow Illumination Model

Here we first revisit the previous illumination formula of modeling shadow removal process. Early traditional methods (Liu and Gleicher 2008; Wu et al. 2007; Arbel and Hel-Or 2010) have been exploring shadow removal based on the simplified physical image formulation, in which shadow images $\mathbf{I}_s$ are modeled as the element-wise multiplication of a shadow matte $\mathbf{S}_m$ and a shadow-free image $\mathbf{I}_{ns}$:

$$\mathbf{I}_s = \mathbf{S}_m * \mathbf{I}_{ns}. \tag{3}$$

This Eqn.(3) is evolved from a universal image formation equation, proposed by (Barrow et al. 1978) as

$$\mathbf{I} = \mathbf{R} * \mathbf{L}, \tag{4}$$

where image $\mathbf{I}$ is obtained by element-wisely multiplying the reflectance $\mathbf{R}$ and luminance $\mathbf{L}$. In (Qu et al. 2017), they further transform Eqn.(3) into log space as

$$\log(\mathbf{I}_s) = \log(\mathbf{S}_m) + \log(\mathbf{I}_{ns}). \tag{5}$$

They directly infer the mapping function between the shadow image $\mathbf{I}_s$ and its matte $\mathbf{S}_m$ via the deep CNNs. However, such simplified Eqn.(3) and (5) do not specifically consider the shadow image characteristics, such as the difference between shadow and non-shadow regions.

(Le and Samaras 2019) firstly proposes a shadow illumination model to separately process shadow/non-shadow regions and fuse them into the second stage as

$$\mathbf{I}_{lit} = \omega * \mathbf{I}_s + b, \tag{6}$$

$$\mathbf{I}_{ns} = \boldsymbol{\alpha} * \mathbf{I}_s + (\mathbf{1} - \boldsymbol{\alpha}) * \mathbf{I}_{lit}. \tag{7}$$

Their modeling function assumes that the shadow effects are completely consistent in the umbra areas, thus they obtain the relit results through the linear transformation with only six constants ($\omega = [\omega_R, \omega_G, \omega_B], b = [b_R, b_G, b_B]$). We argue that applying the consistent transformation for all pixels of the umbra areas is unreasonable. We also prove our view through a simple but effective statistics analysis on the ISTD dataset, as shown in Figure 2. Following (Le and Samaras 2019, 2020), the shadow mask $\mathbf{M}$ can be decomposed into $\mathbf{M}_{umbra}$ and $\mathbf{M}_{penumbra}$ (shadow boundary areas). We utilize pixel erosion operation to obtain $\mathbf{M}_{umbra}$ in Figure 2. Secondly, we could obtain the pixel ratio map of umbra regions $\mathbf{R}$ through

$$\mathbf{R} = \frac{\mathbf{I}_{ns}}{\mathbf{I}_s} * \mathbf{M}_{umbra}. \tag{8}$$

Then we utilize the histogram to analyze the ratio map values that only belong to the umbra area $\mathbf{M}_{umbra}$. We found ratio map histograms of most shadow images are not uniform in the umbra regions. Hence we argue applying uniform transformations in (Le and Samaras 2019, 2020) for all umbra regions pixels is accurate enough.

Figure 2: A simplified representation of one pair shadow images statistical analysis from the ISTD dataset. We find the ratio map R is not uniform distribution in the umbra areas, which shows the shadow image's spatially-variant property in the umbra region. Thus, applying a uniform transformation (*e.g.*, Eqn. (6)) for all pixels in the umbra area is inaccurate. *(Best viewed on-screen.)*

Inspired by the aforementioned models, we propose a new shadow illumination model, which considers the spatially-variant property in umbra areas and exploits existing shadow mask information. The new model can be written as

$$\begin{cases} \mathbf{I}_{ns} = (\mathbf{1} + \mathbf{A}) * \mathbf{I}_s, \\ s.t. \ \ \|(\mathbf{1} - \mathbf{M}) * \mathbf{A})\|_2 = 0, \end{cases} \tag{9}$$

where $\mathbf{M}$ means the shadow mask that shadow regions are 1 and the rest are 0, and $\mathbf{A}$ is the corresponding learned illumination transformation map. Note that we impose a constraint on $\mathbf{A}$, which enforces value of the non-shadow pixel equals its value in the input image and guarantees identical mapping among non-shadow areas. This physically plausible transformation also conforms to the goal of the shadow removal task: recovering shadow regions' information and maintaining the non-shadow regions' information.

Variational Model and Optimization Algorithm

Generally, we formulate this problem as a *maximum a-posteriori* estimation problem, which is written as

$$\begin{aligned} \mathbf{I}_{ns} &= \arg\max_{\mathbf{I}_{ns}} \ \log P(\mathbf{I}_{ns}|\mathbf{I}_s) \\ &= \arg\max_{\mathbf{I}_{ns}} \ \log P(\mathbf{I}_s|\mathbf{I}_{ns}) + \log P(\mathbf{I}_{ns}), \end{aligned} \tag{10}$$

where $\log P(\mathbf{I}_s|\mathbf{I}_{ns})$ indicates the data likelihood and $\log P(\mathbf{I}_{ns})$ characterizes the intrinsic prior knowledge of $\mathbf{I}_{ns}$. Based on our proposed shadow illumination Eqn. (9) and the Eqn. (10), shadow removal problem is transferred to solve the following minimization variational model:

$$\begin{cases} \arg\min_{\mathbf{I}_{ns}} D(\mathbf{I}_s, \mathbf{I}_{ns}, \mathbf{A}) + \lambda\varphi(\mathbf{I}_{ns}, \mathbf{A}), \\ s.t. \qquad g(\mathbf{A}) = 0, \end{cases} \tag{11}$$

where λ is a weight parameter and $D(\cdot)$ denotes the favorable data fidelity term, defined as

$$D(\mathbf{I}_s, \mathbf{I}_{ns}, \mathbf{A}) = \frac{1}{2}\left\|\mathbf{I}_s - \frac{\mathbf{I}_{ns}}{\mathbf{1} + \mathbf{A}}\right\|_2, \tag{12}$$

For the convenience of writing, we employ $g(\mathbf{A})$ to present the constraint term: $\|(\mathbf{1} - \mathbf{M}) * \mathbf{A})\|_2$. We further convert it

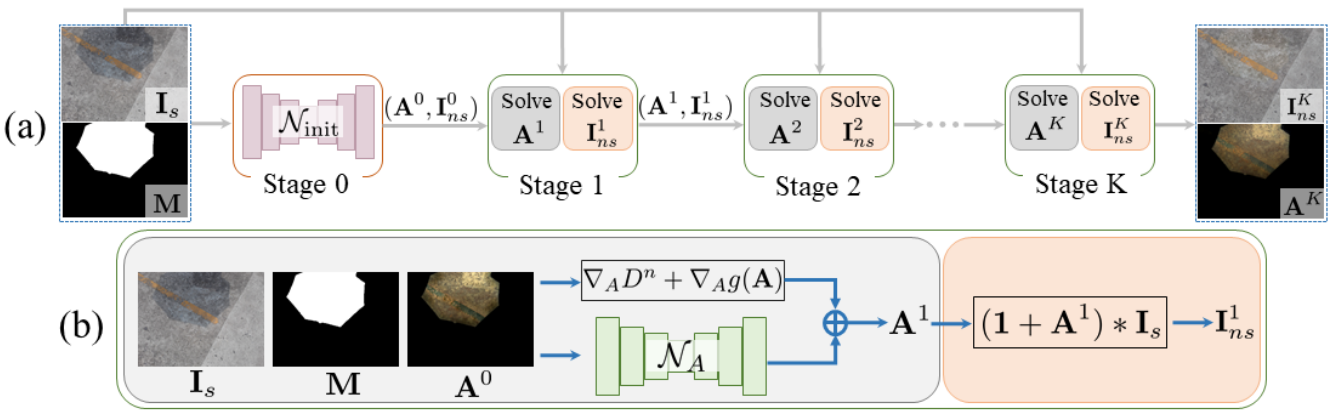

Figure 3: (a) Overview illustration of our proposed efficient model-driven network, which conforms to our designed iterative algorithm 1. (b) The basic structure of one of the iterative stages. Except for the first initialization stage, each subsequent stage consists of two sub-steps, respectively corresponding to Eqn. (14) to update the transformation map $\mathbf{A}$ and shadow-free $\mathbf{I}_{ns}$.

into a non-constrained optimization problem as

$$\arg\min_{\mathbf{I}_{ns}} D(\mathbf{I}_s, \mathbf{I}_{ns}, \mathbf{A}) + \beta g(\mathbf{A}) + \lambda\, \varphi\,(\boldsymbol{A}), \qquad (13)$$

where β is the weight parameter. Since the data fidelity items $D(\cdot)$ and $g(\cdot)$ are both quadratic constraints, they are differentiable. Assume that $\varphi(\cdot)$ is differentiable, the optimization objection (13) can be addressed through the gradient descent scheme (Ruder 2016). At every iteration stage, we need to alternately update $\mathbf{A}$ and $\mathbf{I}_{ns}$ as follows:

$$\begin{cases} \mathbf{A}^{n+1} = \mathbf{A}^n - \eta_A(\nabla_A D^n + \beta \nabla_A g(\mathbf{A}^n) + \lambda \nabla_A \varphi(\mathbf{A}^n)), \\ \mathbf{I}^{n+1}_{ns} = (\mathbf{1} + \mathbf{A}^{n+1}) * \mathbf{I}_s, \end{cases} \qquad (14)$$

where η_A indicates the step size of A; ∇ indicates the gradient operator; n indicates the current iteration number and the maximum iteration number is K.

There are two unresolved problems for the above iteration process in Eqn. (14). One is computing the objective function gradients. The corresponding gradients of the quadratic items $D(\cdot)$ and g$(\cdot)$ are obviously easy to calculate. Followed (Liu et al. 2019), the counterpart of $\varphi(\cdot)$ can be performed with a deep CNNs. $\nabla_A \varphi(\mathbf{A}^n)$ could be directly learned from our training dataset. The other problem is to obtain the initial values for the iteration algorithm. As the CNNs can generate reasonable results, we leverage CNNs technique to estimate the initial results $\mathbf{A}^0$ and $\mathbf{I}^0_{ns}$.

Unfolding Network Design

As shown in Figure 3, we build a deep CNNs model following the proposed iteration algorithm **1**. Specifically, the network contains one initialization stage and K iterative stages, which is corresponding to the total number of iterations of our proposed optimization algorithm. Except for the initial stage, each subsequent stage consists of two sub-steps, respectively corresponding to Eqn. (14) to update the learned transformation map $\mathbf{A}$ and the shadow-free image $\mathbf{I}_{ns}$.

In Figure 3, one place to deploy CNNs is to obtain the initial results, and the other is to solve the gradient result:

$$\begin{aligned} &\mathbf{A}^0, \mathbf{I}^0_{ns} = \mathcal{N}_{init}(\mathbf{I}_s, \mathbf{M}), \\ &\mathcal{N}^{n+1}_A = \nabla_A \varphi(\mathbf{A}^n). \end{aligned} \qquad (15)$$

The network architecture of two CNNs $\mathcal{N}_{init}$ and $\mathcal{N}_A$ both employ classical U-net (Ronneberger, Fischer, and Brox 2015). The network totally involves 4 scales, each of which features a concatenation operation between down-scaling and up-scaling parts. Specifically, the number of channels from the 1^{th} scale to the 4^{th} scale is 32, 64, 128, and 256, respectively. We adopt the 2×2 max pooling operation with stride 2 for the down-scaling operation and 2×2 transposed convolution layer for the up-scaling operation. We further enforce $\mathcal{N}_A$ to share parameters among different stages to decrease the number of network parameters.

Different from the original U-net network, we replace the basic block in each scale. In order to improve our model computation efficiency and save network parameters, we construct two convolution blocks based on the Mobilenet-v2 (Sandler et al. 2018), as shown in Figure 4. The block (b) is derived from ResNet (He et al. 2016), served as the basic block in $\mathcal{N}_A$. While block (c) conforms to our proposed formulation, called the dynamic mapping residual block (DMRB). $\mathcal{N}_{init}$ adopts DMRB as the basic block in each scale.

Network Loss Function

The loss is composed of two parts: the data fidelity term of estimated results; the regularization term of the learned map $\mathbf{A}$. We employ the mean square mean (MSE) to build the network loss function at every iterative stage, defined as:

$$\mathcal{L} = \sum_{n=0}^{K} \gamma^n_d \left\|\mathbf{I}^n_{ns} - \mathbf{I}_{ns}\right\|_2 + \sum_{n=0}^{K} \gamma^n_{reg} \left\|(\mathbf{1} - \mathbf{M}) * \mathbf{A}^n)\right\|_2, \qquad (16)$$

where $\mathbf{I}^n_{ns}$ and $\mathbf{A}^n$ respectively indicate the estimated shadow-free result and the learned transformation map at the n^{th} iteration stage; γ^n_d and γ^n_{reg} indicate tradeoff weights ($\gamma^n_d = 1$ (n = 0,1,$\cdots$,$K-1$),$\gamma^n_K = 10$ and $\gamma^n_{reg} = 1e-6$).

Algorithm 1: Iterative algorithm for shadow removal.

Initialization:
hyperparameters η_A =0.01,β =0.01, λ =0.01; $n \leftarrow 0$;
maximum iterations K = 4; $\mathbf{A}^0, \mathbf{I}_{ns}^0 = \mathcal{N}_{init}(\mathbf{I}_s, \mathbf{M})$;
repeat
  $n \leftarrow n + 1$
  $\mathbf{A}^{n+1} = \mathbf{A}^n - \eta_A(\nabla_A D^n + \beta \nabla_A g(\mathbf{A}^n) + \lambda \mathcal{N}_A^{n+1})$
  $\mathbf{I}_{ns}^{n+1} = (\mathbf{1} + \mathbf{A}^{n+1}) * \mathbf{I}_s$
  Update $\mathcal{N}_{init}$ with Eqn.(16)
until $n + 1 == K$
Output: $\mathbf{I}_{ns}^K$ and $\mathbf{A}^K$

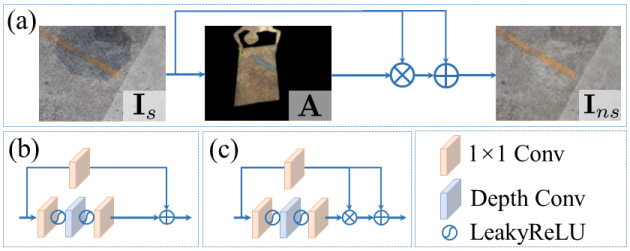

Figure 4: (a) Images illustration of our proposed shadow illumination model (Eqn. (9)); (b) Basic block in $\mathcal{N}_A$, derived from ResNet and Mobilenet-v2; (c) Basic block in $\mathcal{N}_{init}$, a model-informed module from (a).

Experiments

Implementation Details

We implement our network in the PyTorch framework on the PC with a single NVIDIA GeForce GTX 1080Ti GPU. In the training phase, we adopt the Adam optimizer (Kingma and Ba 2014) with a batch size of 2 and the patch size of 256×256. The initial learning rate is 5×10^{-5} and changes with Cosine Annealing scheme (Loshchilov and Hutter 2016). The CNNs parameters are randomly initialized and the model converges well after 150 epochs. For the hyperparameters, the weights (η_A, β, λ) in Equations (14) are initialized as 0.01 and these parameters can be automatically updated during the training phase in an end-to-end manner. The maximum iteration number K is empirically set to 4 as a trade-off between speed and accuracy.

Dataset and Evaluation Metrics

ISTD dataset ISTD is proposed in (Wang, Li, and Yang 2018), which is the first public benchmark that could be used to train shadow detection and removal. ISTD totally consists of 1870 image triplets (shadow image, shadow-free image, shadow mask), including 135 various scenes with different shadow shapes. This dataset has been divided into 1330 triplets for training and 540 triplets for testing.

SRD dataset SRD is proposed in (Qu et al. 2017), containing 2680 and 408 pairs of images for training and testing respectively. SRD does not provide masks, we employ the public SRD shadow masks from (Cun, Pun, and Shi 2020).

For evaluation on ISTD and SRD datasets, we compute the root mean square error (RMSE) between the estimated result and ground truth image in the LAB color space. For fair comparisons, we directly adopt the MATLAB evaluation codes from (Fu et al. 2021a). Following previous methods (Liu et al. 2021c; Fu et al. 2021a; Le and Samaras 2020, 2019), we employ results with a resolution of 256 × 256 for evaluation in this paper. In the Table 1 and 2, we report the evaluation results about the shadow (S) regions, non-shadow (NS) regions, and all images (ALL). For the RMSE metric, the lower value means the better result. In addition, we also compute the PSNR and SSIM (Wang et al. 2004) for quantitative assessment in RGB color space.

Shadow Removal Evaluation on ISTD Dataset

We first compare our ISTD results with current state-of-the-art shadow removal methods, including 2 traditional methods (Guo, Dai, and Hoiem 2012; Gong and Cosker 2014) and 9 CNNs-based methods (Wang, Li, and Yang 2018; Hu et al. 2019a,b; Le and Samaras 2019, 2020; Cun, Pun, and Shi 2020; Liu et al. 2021b,c; Fu et al. 2021a). For fair comparison, all the comparison results or metrics values are provided by the original authors. Since the pre-trained model and shadow removal results of MaskShadow-GAN (Hu et al. 2019b) are not publicly available, we directly adopt the metrics values from (Fu et al. 2021a). Quantitative comparisons are shown in Tabel 1. We surpass the SOTA methods on the PSNR metric by a large margin, whether in the shadow region, non-shadow region, or the entire region. Additionally, we compare visual results in the Figure 5, in which our method achieves better visual performance than other methods. (Le and Samaras 2020; Fu et al. 2021a) may misprocess relatively dark non-shadow regions, bringing undesired artifacts in their estimated results. It turns out that their models fail to make full use of the shadow mask information, even though their network inputs contain shadow masks.

We further compare the different models efficiency and complexity on ISTD results, as shown in Figure 1 and Table 6. Our method achieves leading shadow removal performance in terms of quantitative metrics, inference efficiency, and model interpretability. We also evaluate our model on the AISTD dataset (Le and Samaras 2019). Our PSNR ↑ and RMSE ↓ are 33.36 dB and 3.42, while metric values of SOTA method (Fu et al. 2021a) are 29.44 dB and 4.23.

Shadow Removal Evaluation on SRD Dataset

We also compare the methods (Qu et al. 2017; Hu et al. 2019a; Cun, Pun, and Shi 2020; Fu et al. 2021a), which provide the pre-trained model or processed results on the SRD dataset. During the evaluation, we adopt the public SRD shadow masks provided by (Cun, Pun, and Shi 2020). The comparison results can be found in Table 2. It is obvious that our method achieves the best performance on all metrics.

Ablation Study

Ablation Study of Iterative Stage Numbers The effectiveness of the iteration numbers is tested by comparing the performance on the PSNR and RMSE metrics. Table 5 demonstrates the effect of the stage number K on the shadow removal performance of our network. $K = 0$ represents

Method	Venue	Shadow Region (S)			Non-Shadow Region (NS)			All image (ALL)		
		PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
Original input	-	22.40	0.9361	32.10	27.32	0.9755	7.09	20.56	0.8934	10.88
Guo *et al.*	PAMI'12	27.76	0.9643	18.65	26.44	0.9664	7.76	23.08	0.9198	9.26
Gong *et al.*	BMVC'14	30.14	0.9727	13.54	26.98	0.9730	7.20	24.71	0.9260	8.03
ST-CGAN	CVPR' 18	33.74	0.9808	9.99	29.51	0.9576	6.05	27.44	0.9291	6.65
MaskShadow-GAN‡	ICCV' 19	-	-	12.67	-	-	6.68	-	-	7.41
SP-M-Net	ICCV' 19	32.16	0.9812	10.30	26.40	0.9702	7.47	25.08	0.9429	7.79
DSC	T-PAMI' 19	34.64	0.9835	8.72	31.26	0.9690	5.04	29.00	0.9438	5.59
Param+M+D-Net	ECCV' 20	31.43	0.9811	11.84	26.21	0.9687	7.51	24.69	0.9413	7.94
DHAN	AAAI' 20	35.53	**0.9882**	**7.73**	31.05	0.9705	5.29	29.11	0.9543	5.66
LG-ShadowNet	T-IP' 21	30.88	0.9794	11.32	25.42	0.9639	8.01	23.89	0.9340	8.37
G2R	CVPR' 21	31.63	0.9746	10.72	26.19	0.9671	7.55	24.72	0.9324	7.85
Auto-Exposure	CVPR' 21	34.71	0.9752	7.91	28.61	0.8799	5.51	27.19	0.8456	5.88
Ours	-	**36.95**	0.9867	8.29	**31.54**	**0.9779**	**4.55**	**29.85**	**0.9598**	**5.09**

Table 1: Quantitative comparisons of the SOTA methods on the ISTD datasets. The best and the second results are boldfaced and underlined, respectively. ‡ means results copied from Auto-Exposure (Fu et al. 2021a).

Method	Venue	Shadow Region (S)			Non-Shadow Region (NS)			All image (ALL)		
		PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
Original input	-	18.96	0.8710	36.69	31.47	0.9750	4.83	18.19	0.8295	14.05
Guo *et al.*‡	PAMI' 12	-	-	29.89	-	-	6.47	-	-	12.60
DeshadowNet‡	CVPR' 17	-	-	11.78	-	-	4.84	-	-	6.64
DSC	T-PAMI' 19	30.65	0.9602	8.62	31.94	0.9650	4.41	27.76	0.9033	5.72
DHAN	AAAI' 20	33.67	0.9777	**7.16**	34.79	0.9789	3.91	30.51	0.9492	4.87
Auto-Exposure	CVPR' 21	32.26	0.9663	8.55	30.59	0.9445	5.74	27.74	0.8933	6.50
Ours	-	**34.94**	**0.9797**	7.44	**35.85**	**0.9819**	**3.74**	**31.72**	**0.9523**	**4.79**

Table 2: Quantitative comparisons of the SOTA methods on the SRD datasets. The best and the second results are boldfaced and underlined, respectively. ‡ means results copied from Auto-Exposure (Fu et al. 2021a).

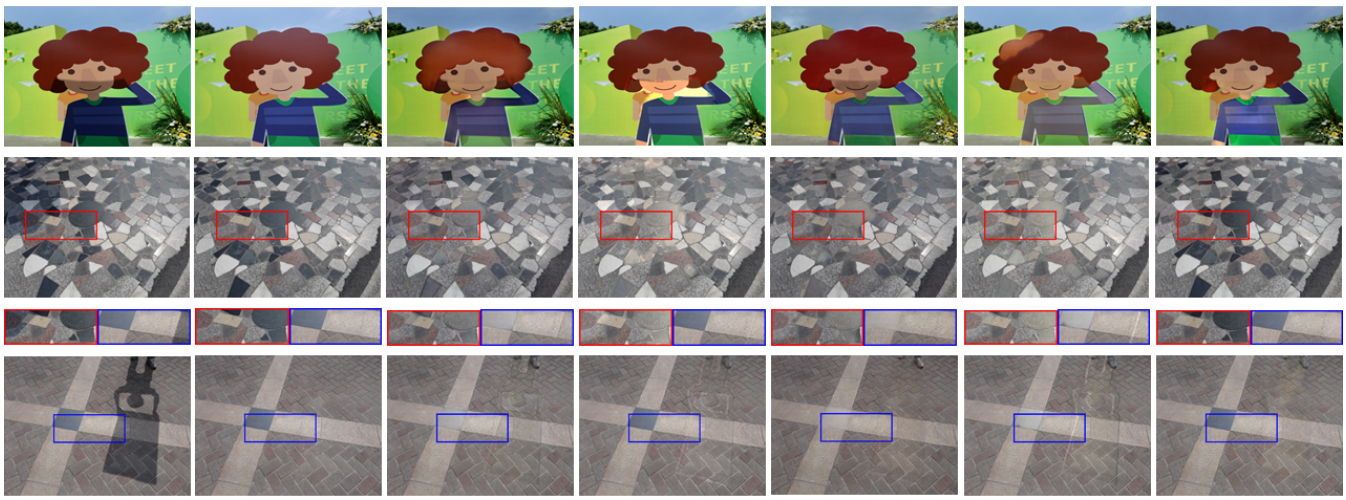

Figure 5: Visual comparisons with SOTA methods on ISTD datset. (a) to (d) are the estimated results from SOTA methods : DSC (Hu et al. 2019a), Param+M+D-Net (Le and Samaras 2020), DHAN (Cun, Pun, and Shi 2020), and G2R (Liu et al. 2021c).

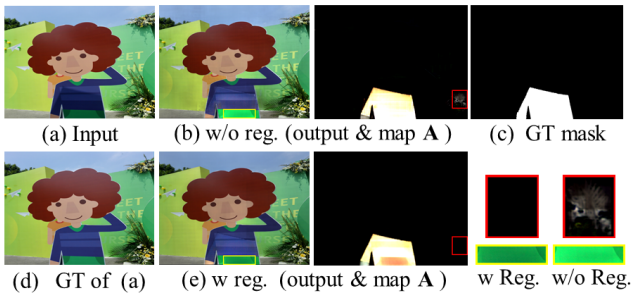

Figure 6: The visual comparison of results whether the loss function contains regularization term.

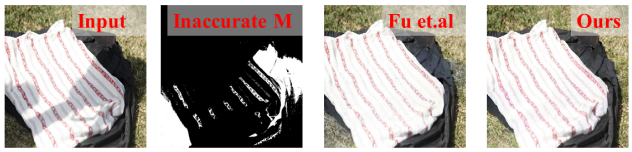

Figure 7: The visual comparisons with imperfect mask.

Models	Initial Network		w Iterations	All image (ALL)	
	RB	DMRB		PSNR↑	RMSE↓
Model-1	✓			28.37	5.74
Model-2		✓		28.52	5.50
Model-3	✓		✓	29.44	5.17
Ours		✓	✓	**29.85**	**5.09**

Table 3: Ablation study of our proposed DMRB module.

Models	S		NS		ALL	
	PSNR↑	RMSE↓	PSNR↑	RMSE↓	PSNR↑	RMSE↓
w/o reg.	35.86	9.23	31.19	4.81	29.41	5.46
Ours	**36.95**	**8.29**	**31.54**	**4.55**	**29.85**	**5.09**

Table 4: Ablation study of the regularization term in loss.

the results estimated by $\mathcal{N}_{init}$, which does not include any subsequent iteration stages. Taking $K = 0$ as the baseline, it can be seen that the shadow removal performance of our method gradually increased to a peak at $K = 4$. Meanwhile, it also reaches the lowest value at $K = 4$ for the RMSE. Therefore, we empirically set $K = 4$ as the default setting.

Ablation Study of DMRB Module In this ablation study, we compare the model performance of using Resblock (R-B) or Dynamic Mapping Residual Block (DMRB) as basic blocks in the initial network $\mathcal{N}_{init}$. $\mathcal{N}_{init}$ provides initial values ($\mathbf{A}^0$, $\mathbf{I}^0_{ns}$) for the subsequent iterative algorithm. Moreover, the relationship between $\mathbf{A}^0$ and $\mathbf{I}^0_{ns}$ conforms to our proposed shadow illumination model. Intuitively, we enforce the network intermediate features extraction following the proposed shadow model. According to Table 3, the model using DMRB obviously performs better than using RB under the same setting. This is because DMRB integrates shadow model-self knowledge that benefits shadow removal. Extensive experiments indicate DMRB boosts shadow removal performance, without introducing additional parameters.

Models	Iterative Stage Numbers					
	0	1	2	3	4	5
PSNR ↑	28.52	28.85	28.98	29.32	**29.85**	29.70
RMSE ↓	5.50	5.55	5.36	5.19	**5.09**	5.18

Table 5: Ablation study of iterative stage numbers.

Method	DHAN	SP-M-Net	G2R	Ours
Inference Time	0.3450s	0.0376s	0.0176s	0.0197s

Table 6: Average inference time on the 1080Ti GPU device with the resolution of 480 × 640.

Ablation Study of Regularization Term We firstly verify the effect of the regularization term of the learned transformation map **A** in the loss function. Based on our proposed shadow illumination model, the ideal transformation is to process only the shadow regions while maintaining the identity mapping of the non-shaded area. From Figure 6, adding the regularization term could reduce the unwanted artifacts on transformation map **A**. In Table 4, the performance of shadow removal is obviously decreased without the regularization term.

We find that our network is not sensitive to the value of γ^n_{reg}, as long as the range of γ^n_{reg} is between $1e-5$ and $1e-7$. Therefore, we empirically set $\gamma^n_{reg} = 1e-6$ (n = 0,1,$\cdots$,K) as the default setting throughout all experiments.

Ablation Study of the Imperfect Mask In Figure 7, we provide a visualization comparison with the imperfect shadow mask as input. Our method still could remove the shadows and recover the underlying contents.

Ablation Study of Inference Time In Table 6, we compare part SOTA methods' average inference time on ISTD testing dataset. Our method's inference speed is over 17 times faster than DHAN (Cun, Pun, and Shi 2020).

Conclusion

In our paper, we verify the previous shadow illumination models exists drawbacks, *e.g.*, ignoring the spatially-variant property of shadow images. We further propose a new shadow illumination model, which considers spatially-variant property and shadow mask information. Utilizing the new shadow illumination model, we reformulate the shadow removal task as a variational optimization problem. To this end, we build an iterative optimization algorithm to address it. This iteration optimization algorithm is unrolled into deep CNNs, so that our method enjoys the advantages of both model-driven methods (*e.g.*, well interpretability) and data-driven CNNs-based methods (*e.g.*, powerful learning capability). Extensive experiments demonstrate that our method achieves leading shadow removal performance with fewer network parameters and faster inference speed.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2020AAA0105702, National Natural Science Foundation of China (NSFC) under Grants U19B2038 and 61901433, the University Synergy Innovation Program of Anhui Province under Grants GXXT-2019-025, the USTC Research Funds of the Double First-Class Initiative under Grant YD2100002003.

References

Afonso, M. V.; Bioucas-Dias, J. M.; and Figueiredo, M. A. T. 2010. Fast Image Recovery Using Variable Splitting and Constrained Optimization. *TIP*.

Arbel, E.; and Hel-Or, H. 2010. Shadow removal using intensity surfaces and texture anchor points. *PAMI*.

Barrow, H.; Tenenbaum, J.; Hanson, A.; and Riseman, E. 1978. Recovering intrinsic scene characteristics. *Comput. Vis. Syst*, 2.

Beck, A.; and Teboulle, M. 2009. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM*.

Cucchiara, R.; Grana, C.; Piccardi, M.; and Prati, A. 2003. Detecting moving objects, ghosts, and shadows in video streams. *PAMI*, 25(10): 1337–1342.

Cun, X.; Pun, C.-M.; and Shi, C. 2020. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN. In *AAAI*.

Ding, B.; Long, C.; Zhang, L.; and Xiao, C. 2019. ARGAN: Attentive Recurrent Generative Adversarial Network for Shadow Detection and Removal. In *ICCV*.

Dong, W.; Wang, P.; Yin, W.; Shi, G.; Wu, F.; and Lu, X. 2018. Denoising prior driven deep neural network for image restoration. *PAMI*.

Fattal, R. 2008. Single image dehazing. *TOG*.

Finlayson, G. D.; Hordley, S. D.; and Drew, M. S. 2002. Removing shadows from images. In *ECCV*.

Fu, L.; Zhou, C.; Guo, Q.; Juefei-Xu, F.; Yu, H.; Feng, W.; Liu, Y.; and Wang, S. 2021a. Auto-exposure fusion for single-image shadow removal. In *CVPR*.

Fu, X.; Wang, M.; Cao, X.; Ding, X.; and Zha, Z.-J. 2021b. A Model-Driven Deep Unfolding Method for JPEG Artifacts Removal. *TNNLS*.

Fu, X.; Zha, Z.-J.; Wu, F.; Ding, X.; and Paisley, J. 2019. JPEG Artifacts Reduction via Deep Convolutional Sparse Coding. In *ICCV*.

Gong, H.; and Cosker, D. 2014. Interactive Shadow Removal and Ground Truth for Variable Scene Categories. In *BMVC*.

Guo, Q.; Xie, X.; Juefei-Xu, F.; Ma, L.; Li, Z.; Xue, W.; Feng, W.; and Liu, Y. 2020. SPARK: Spatial-aware Online Incremental Attack Against Visual Tracking. In *ECCV*.

Guo, R.; Dai, Q.; and Hoiem, D. 2012. Paired regions for shadow detection and removal. *PAMI*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hu, X.; Fu, C.-W.; Zhu, L.; Qin, J.; and Heng, P.-A. 2019a. Direction-aware Spatial Context Features for Shadow Detection and Removal. *PAMI*.

Hu, X.; Jiang, Y.; Fu, C.-W.; and Heng, P.-A. 2019b. Mask-ShadowGAN: Learning to Remove Shadows from Unpaired Data. In *ICCV*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Le, H.; and Samaras, D. 2019. Shadow Removal via Shadow Image Decomposition. In *ICCV*.

Le, H.; and Samaras, D. 2020. From shadow segmentation to shadow removal. In *ECCV*.

Li, L.; Pan, J.; Lai, W.-S.; Gao, C.; Sang, N.; and Yang, M.-H. 2019. Blind image deblurring via deep discriminative priors. *IJCV*.

Liu, F.; and Gleicher, M. 2008. Texture-consistent shadow removal. In *ECCV*.

Liu, R.; Ma, L.; Zhang, J.; Fan, X.; and Luo, Z. 2021a. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *ICCV*.

Liu, Y.; Pan, J.; Ren, J.; and Su, Z. 2019. Learning deep priors for image dehazing. In *ICCV*.

Liu, Z.; Yin, H.; Mi, Y.; Pu, M.; and Wang, S. 2021b. Shadow Removal by a Lightness-Guided Network with Training on Unpaired Data. *TIP*.

Liu, Z.; Yin, H.; Wu, X.; Wu, Z.; Mi, Y.; and Wang, S. 2021c. From Shadow Generation to Shadow Removal. In *CVPR*.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Mu, P.; Chen, J.; Liu, R.; Fan, X.; and Luo, Z. 2018. Learning bilevel layer priors for single image rain streaks removal. *SPL*.

Narasimhan, S. G.; and Nayar, S. K. 2000. Chromatic framework for vision in bad weather. In *CVPR*.

Qu, L.; Tian, J.; He, S.; Tang, Y.; and Lau, R. W. 2017. Deshadownet: A multi-context embedding deep network for shadow removal. In *CVPR*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer.

Ruder, S. 2016. An overview of gradient descent optimization algorithms. *arXiv*.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*.

Sanin, A.; Sanderson, C.; and Lovell, B. C. 2010. Improved shadow removal for robust person tracking in surveillance scenarios. In *ICPR*.

Shor, Y.; and Lischinski, D. 2008. The shadow meets the mask: Pyramid-based shadow removal. In *CGF*.

Wang, H.; Xie, Q.; Zhao, Q.; and Meng, D. 2020. A model-driven deep neural network for single image rain removal. In *CVPR*.

Wang, J.; Li, X.; and Yang, J. 2018. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *TIP*.

Wu, Q.; Zhang, W.; and Kumar, B. V. 2012. Strong shadow removal via patch-based shadow edge detection. In *ICRA*.

Wu, T.-P.; Tang, C.-K.; Brown, M. S.; and Shum, H.-Y. 2007. Natural shadow matting. *TOG*.

Zhang, K.; Gool, L. V.; and Timofte, R. 2020. Deep unfolding network for image super-resolution. In *CVPR*.

Zhang, K.; Zuo, W.; Gu, S.; and Zhang, L. 2017. Learning deep CNN denoiser prior for image restoration. In *CVPR*.