This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Physics Based Image Deshadowing Using Local Linear Model

Tamir Einy[1*] Efrat Immer[2*] Gilad Vered[1] Shai Avidan[2]

Applied Materials[1] Tel-Aviv University[2]
{tamireiny, efratimmer, vgilad}@gmail.com, avidan@eng.tau.ac.il

Abstract

Image deshadowing algorithms remove shadows from images. This requires both detecting where the shadow is and, once detected, removing it from the image. This work focuses on the shadow removal part. We follow a common physical shadow formation model and learn its parameters using a deep neural network. Our model consists of an existing network for shadow detection, and a novel network for shadow removal. The shadow removal network gets the predicted mask of the shadow region and the shadow image and predicts six parameters per pixel. Remarkably, a straightforward network architecture, that is considerably smaller compared to alternative methods, produces better results on standard datasets[1].

1. Introduction

Shadows follow us wherever we go. Virtually every image we capture contains shadows that impact a variety of Computer Vision applications such as object detection, recognition and image segmentation. This is why detecting and removing shadows can prove useful.

Classic methods use physical image formation models to deshadow an image. The parameters of the light source and the surface material are estimated in the case of no occlusions, and used to produce a shadow-free image. In the common physical model there are 6 unknowns per pixel (gain and bias per channel) and some prior, or user assistance is required in order to solve the problem.

This changed with the rise of deep learning models, and the release of large scale datasets. It is now possible to train deep neural networks to estimate these unknowns, without making explicit prior assumptions or resorting to human intervention.

Various Deep Learning algorithms have been proposed to solve the problem using different architectures and estimating various types of parameters. One approach treats image deshadowing as an image-translation problem where the goal is to translate a shadow image to a shadow-free image. However, in practice we found that this creates somewhat blurry images and instead opt for a physics based approach. In this approach, the goal is to estimate the parameters of the physical model. Once the parameters are recovered we can use them to recover the shadow-free image.

Image deshadowing can be separated into shadow detection and shadow removal. A shadow detection algorithm takes an input image and produces a mask (either binary or probabilistic) that determines which pixels are affected by the shadow, and to what extent. We build upon recent work and use an existing shadow detection network to detect the shadow mask.

Our shadow removal part takes the input image and its corresponding mask and removes the shadows accordingly. To do that, we suggest a simple network that directly estimates 6 parameters per pixel. The resulting network is considerable smaller, in terms of the number of its parameters, than what was previously reported in the literature, and we obtain an improved model that achieves better results on standard datasets.

To summarize, the main contributions of our paper are:

- A novel network for estimating a physical model that consists of 6 parameters per pixel
- Smaller network for shadow removal (100k parameters) vs. 100M for other methods
- State of the art deshadowing performance on standard datasets

2. Related work

Physics based methods. Early works often utilized a physical shadow formation model. Barrow and Tenenbaum [2] introduced the notion of "intrinsic images" to represent the idea of decomposing an image into reflectance and illumination components. The shadow-free image can be recovered by extracting the changes in the image which arise from changes in the scene illumination. Different

*Denotes equal contribution

[1]Code is available at https://github.com/tamireiny/local_linear_deshadow

shadow-formation models have been suggested to recover the intrinsic images.

Finlayson *et al*. [5] created a shadow-invariant image by projecting the image colors onto a direction orthogonal to that of the illumination change. The invariant image is used to detect the shadow region and the shadows are removed by zeroing the gradients on the shadow edge.

Shor *et al*. [16] derived an affine shadow formation model by expressing the illumination as the sum of ambient and direct illuminations and describing the shadow formation by blocking completely the direct light and partially the ambient light. They estimate 4 parameters to model the affine relationship between the shadow and the shadow-free regions. The shadow region is divided into several areas to account for shadow variations.

Arbel and Hel-Or [1] pointed out some of the challenges related to real-world shadow removal task. For example, shadows can be non-uniform, i.e., the shadow intensity or color varies within the shadow region. This is caused due to interactions between the ambient light and the occluding object or by the geometry of the shadowed surface. To address this challenge they estimate a per-pixel per-channel shadow scale factor that allows a spatially varying correction. The scale factor is calculated by fitting a smooth intensity surface to the approximated shadow-free region using a thin-plate surface model.

Deep-learning based methods. More recently, the rise of learning-based methods and the publication of large-scale shadow removal datasets such as SRD [15] and ISTD [18] have led to a significant improvement in shadow removal performance. These methods can be divided into two categories: (i) methods that utilize a physical model for shadow formation, training a network to estimate the model parameters, and (ii) methods that treat the problem as an image-to-image translation and disregard the physical model.

The first category includes the works of Le and Samaras [11, 12], Qu *et al* [15] and Fu *et al.* [6]. Qu *et al* [15] proposed the Deshadow-Net which directly learns the shadow matte layer that represents the illumination attenuation caused by the shadow. The shadow-free image is recovered by applying the shadow matte weights to the shadow image. Their framework extracts multi-context features, involving semantics and appearance information which are then used to predict the shadow matte layer.

Le and Samaras [11] created the SP+M-net that is based on a physical model of shadow formation. They formulated a linear shadow formation model in which the shadow image can be expressed as a function of the shadow-free image, the shadow parameters, and a matting mask. The first network, named SP-net, predicts a single set of the shadow parameters using a regression loss with a pre-calculated shadow parameters set (calculated using LS regression between the shadow and the shadow-free images). The second network, named M-net, estimates a matting mask. The shadow parameters are used to relight the shadow image and the matting mask is used to combine the relit image with the shadow image. In [12], Le and Samaras presented a patch-based semi-supervised network. By incorporating an adversarial framework with their physical model, they were able to train the network on unpaired data.

Fu *et al.* [6] formulated the shadow removal as an exposure fusion problem. Their framework predicts exposure levels to create multiple over-exposed images and then a second network computes pixel-wise kernels for fusing the original shadow image with the over-exposed versions.

The second category, that treats image deshadowing as an image-translation problem, includes the works of Wang *et al.* [18], Hu *et al.* [9], Cun *et al.* [4], Hu *et al.* [8] and Liu *et al.* [14].

The ST-CGAN of Wang *et al.* [18] employed a stacked conditional GAN framework for joint shadow detection and removal. The first generator produces a shadow detection mask while the second generator directly predicts the shadow-free image. Hu *et al.* developed the MaskShadow-GAN [9], a variant of the cycle-GAN that is trained on unpaired shadow and shadow-free images. The generator of the shadow image also gets the shadow mask image to produce a shadow image.

DHAN proposed by Cun *et al.* [4] predicts the shadow mask and the shadow-free image in an end-to-end manner. Their network is based on the context aggregation network [3] with additional hierarchical aggregation of multi-contexts features and attentions to better learn the shadow region. They also designed a shadow matting GAN to synthesize realistic shadow images from a given shadow mask and shadow-free image and used it to enhance the train data.

Hu *et al.* [8] design DSC, a direction-aware spatial context module that includes spatial RNNs. They embedded multiple copies of DSC modules in a convolutional neural network to learn features in different scales. The same framework can be trained for shadow detection or for shadow removal by replacing the shadow masks with the shadow-free images as the ground truth and changing the loss.

G2R-ShadowNet of Liu *et al.* [14] is a GAN based weakly supervised network that generates pseudo shadows in the shadow-free region taking the shadow image and shadow mask as input. The generated images are used to train the shadow-removal and shadow-refinement sub-networks.

Shadow detection. Zhu *et al.* [22] use the architecture of Xie *et al.* [19] to extract image features. Spatial information from any two adjacent feature maps is extracted using Recurrent Attention Residual (RAR) Modules. The network is bi-directional, using both top-down and bottom-up ap-

proaches. The output is generated by combining both paths.

Zheng *et al.* [21] focus on the false negative and false positive regions of the shadow binary mask by distraction-aware shadow module. They proposed a network that focuses on the hard areas of shadow image. Specifically, they use this module at several resolution levels and fuse between the modules outputs.

A different approach was proposed by Le *et al.* [13]. Since it is laborious to collect large amount of shadow images and their corresponding masks, a GAN was used to annotate the shadow-affected images. This is done by attenuating the shadow area in the image, according to the labeled mask. Shadow region in those annotations is harder to detect, thus improving the performance of the detector.

Earlier works such as Gong and Cosker [7] and Shor *et al* [16] use semi-interactive methods for the shadow detection that requires the user input.

While some of the shadow removal works [11, 12, 14] made use of a dedicated shadow detection network prior to their shadow removal step, other works [4, 8, 15, 18] performed both the shadow removal and detection jointly in an end-to-end manner.

3. Shadow formation model

In this section we describe a common physical shadow formation model and the assumptions that were previously taken to simplify it. We then examine the validity of this model on real-world shadow images and observe that under the previously taken assumptions, this model is not sufficient to represent the shadow formation in real scenes. To improve this model, we propose to extend it to a local linear shadow model where the shadow parameters can vary across the shadow region.

3.1. Linear shadow model

We use the physical shadow formation model described in previous works [11, 16]. The purpose of this model is to map the relationship between a shadowed pixel to its corresponding lit pixel. The light intensity reflected from a point on a diffusing surface depends on the reflectance of the surface material and the scene illumination. For a point lying on the shadow-free region, the illumination can be expressed as the sum of direct and ambient illumination components. Thus, the light intensity reflected from the shadow-free region is:

$$I_x^{shadow-free}(\lambda) = R_x(\lambda)\left(L_x^d(\lambda) + L_x^a(\lambda)\right) \quad (1)$$

where $I_x^{shadow-free}(\lambda)$ is the light intensity reflected from point x in the scene at wavelength λ, $R_x(\lambda)$ is the reflectance, L_x^d is the direct illumination and L_x^a is the ambient illumination.

The shadow region can be partitioned into the umbra and penumbra areas. The umbra area is fully shadowed while the penumbra area is a transitional area at the outer boundary of the shadow that is partially shadowed. In the umbra area, the direct illumination is completely occluded and the ambient light is partially occluded. Thus, the light intensity reflected from the umbra area is:

$$I_x^{shadow}(\lambda) = a_x(\lambda)R_x(\lambda)L_x^a(\lambda) \quad (2)$$

where $a_x(\lambda)$ is the attenuation factor indicating the remaining fraction of the ambient illumination that arrives at point x in wavelength λ. From Eq. 1 and 2 the relationship between the intensity reflected from the umbra and shadow-free areas can be derived:

$$I_x^{shadow-free}(\lambda) = \underbrace{a_x(\lambda)^{-1}}_{w_x(\lambda)} I_x^{shadow}(\lambda) + \underbrace{L_x^d(\lambda)R_x(\lambda)}_{b_x(\lambda)} \quad (3)$$

where $w_x(\lambda)$ and $b_x(\lambda)$ are used to denote the linear model coefficients.

The model of image acquisition by a camera that relates the scene reflected intensity to the actual pixel intensity is:

$$I_i(k) = \int I_x(\lambda)S^k(\lambda)d\lambda \quad (4)$$

where $I_i(k)$ is the grey level of pixel i and color channel k $k \in \{R, G, B\}$ and $S^k(\lambda)$ is the camera sensor sensitivity for color channel k. Assuming that $a_x(\lambda)$ does not change rapidly across the sensor's spectral range and that the color acquisition process of the camera is linear as shown in Eq. 4, we get the following linear equation:

$$I_i^{shadow-free}(k) = w_i(k)I_i^{shadow}(k) + b_i(k) \quad (5)$$

By calculating the values of $w_i(k)$ and $b_i(k)$, i.e. the shadow parameters, and applying them to the shadow image, the shadowed area in the image can be removed.

In its most general form, this model has 6 unknowns per pixel which leads to a total of $6n$ parameters, where n is number of pixels in the shadow region. To simplify the model, [16] further assumed that $a_x(\lambda)$ is constant for the entire spectral range of the camera and thus $w_i(k)$ does not depend on the color channel k. This resulted in 4 unknowns per pixel. Furthermore, they divided the shadow region into smaller areas and the same set of shadow parameters was calculated for each area. In [11] they assumed that $w_i(k)$ and $b_i(k)$ are both dependent on the color channel k, but are constant across all pixels in the umbra area. Thus, a single set of shadow parameters $w(k), b(k)$ is calculated and used to relit the shadow image:

$$I_i^{relit}(k) = w(k)I_i^{shadow}(k) + b(k) \quad (6)$$

To account for variations in the shadow region, specifically in the penumbra area, [11] estimated an additional

matting layer that is applied to linearly combine the shadow image with the relight image:

$$\hat{I}_i^{shadow-free}(k) = \alpha_i(k) I_i^{shadow}(k) + (1-\alpha_i(k)) I_i^{relit}(k) \quad (7)$$

The matting layer $\alpha_i(k)$ adds one unknown parameter per channel per pixel, resulting in a total of $3n + 6$ unknown parameters.

When examining real-world images, we find that the linear model with the previously taken assumptions is not sufficient for describing the relationship between the shadow and shadow-free pixels. This happens for various reasons. For one, shadows are not uniform. For example, the attenuation of the ambient light can vary both in intensity and in spectral distribution across the shadowed region due to interactions between the illumination and the occluding objects. Another reason is that the geometry of the shadow surface may not be flat. Finally, as can be seen in Eq. 3, the bias term of the shadow parameters depends on the pixel reflectance. These reasons motivate us to explore a more general linear model.

3.2. Local linear shadow model

We propose to use the generalized form of the shadow formation model that estimates w, b per pixel, per channel. The advantage of this form is that it allows the shadow parameters to vary throughout the shadow region. We find that using a small network with only $100k$ parameters (see table 4 for comparison with other methods) serves mainly as a regularizer and leads to piecewise smooth coefficient maps (see example in Fig 1 and in the supplemental). The learnt coefficient maps are applied to the shadow image to recover the shadow-free image as in Eq. 5.

4. Method

4.1. Network architecture

Our pipeline is divided into two main parts: shadow detection and shadow removal, which are implemented using two different neural networks. The shadow detection network input is a shadow image, and it predicts either a binary or probabilistic shadow mask.

The shadow removal network is shown in Fig 1. It takes as input a shadow image and an optional mask (binary or probabilistic) and outputs 6 channels with the same resolution of the shadow image ((w, b) per pixel, per channel). Later, using Eq. 5, we predict the shadow-free image.

Shadow Detection Network Architecture We use the network architecture and initial weights suggested by Zhu *et al*. [22], this network is referred as BDRAR.

Shadow Removal Network Architecture We use the multi-scale context aggregation network (CAN), developed in the context of semantic image analysis. The input image and output have the same resolution. We are using the architecture suggested by Chen *et al*. [3], which includes 10 layers. The first 8 layers contain convolutions with increasing dilation step size. The 9$^{\text{th}}$ layer is convolution layer without dilation. For these 9 layers we use the leaky ReLU activation function presented by Xu *et al*. [20], for the last layer we use a linear transformation (1×1 convolution) without non-linearity, and change its output to be 6 channels, instead of 3 (see network architecture in the supplemental). The usage of dilation allows the network to process information from large receptive field (513×513), while using small number of parameters. Our network minimizes the L_2 loss function:

$$loss = \sum_{k=1}^{3} \sum_{i=1}^{N} (\hat{I}(k)_i^{shad.-free} - I(k)_i^{shad.-free})^2 \quad (8)$$

Applying Eq. 5 in Eq. 8 we get:

$$loss = \sum_{k=1}^{3} \sum_{i=1}^{N} (\hat{w_i}(k) \cdot I(k)_i^{shad.} + \hat{b_i}(k) - I(k)_i^{shad.-free})^2 \quad (9)$$

4.2. Implementation Details

Shadow Detection Network: To generate the binary or probabilistic masks we fine-tuned BDRAR on ISTD+ or SRD datasets. As initial weights, we used the weights published by Zhu *et al*. [22] that trained the model on the SBU dataset [17] for 3000 iterations. The model outputs probability masks, while the binary masks were created by applying CRF [10] refinement followed by thresholding. We split the ISTD+ and SRD train sets to train and validation sets, with approximately 5% of the original train set serving as the validation set. We split the original ISTD+ train set such that the same scene would not appear both in the train and validation sets.

We evaluate the results with IoU and BER (balance error rate). Using fine-tuning, we managed to increase the mean IoU, on the ISTD+ test set, from 0.794 to 0.91 (see example in supplemental). At the same time, the BER dropped from 5.61 to 1.94. We obtain similar improvements on the SRD dataset.

Shadow Removal Network: The network predicts w, b maps from the shadow image and the predicted shadow mask. We then use Eq. 5 to predict the shadow-free image. We further clip the result to be within valid image range of $[0, 255]$. The network loss function is described in Eq. 9. The network weights are initialized with the identity matrix. The training runs for 500 epochs with constant

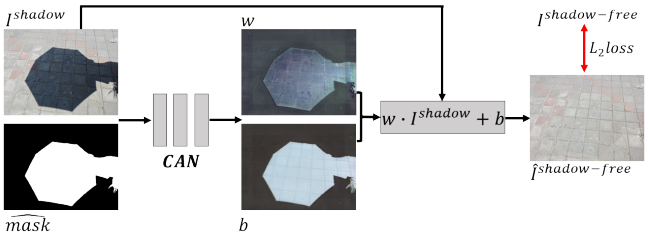

Figure 1. **System architecture:** Our de-shadow system contains two parts: detection and removal. The detection part (not seen in the figure) takes as input a shadow image and outputs the predicted shadow mask. Our removal part gets the predicted mask and the shadow image and predicts w, b maps, each map has the same dimensions as the shadow image. The predicted shadow-free image is calculated pixel wise using Eq. 5.

learning rate of $1 \cdot e^{-4}$. During training an augmentation of random flip (vertical, horizontal or both) is used. All the train set is used for training, as we do not use a validation set. The last model is considered the best and is used for inference and evaluation.

5. Experiments

5.1. Datasets and evaluation metric

ISTD & ISTD+: ISTD, which was created by Wang *et al*. [18], contains triplets of shadow, shadow free and binary mask images captured from 135 different scenes. In total there are 1870 triplets, where 1330 are part of the train set and 540 are part of the test set. Since the dataset was collected outdoors without any lighting control, the non-shadow area in the shadow-free image and shadow image do not necessarily match. The difference was noted by Le and Samaras [11] that suggested a color correction method, using linear regression, to reduce the effect. The revised dataset is denoted as ISTD+.

The ISTD dataset contains images of size 640×480, so a network with a smaller receptive field can be used. Therefore, we remove the last dilating layer and use a network with 9 layers. We use 32 channels per layer resulting in a network with a total number of only $67K$ trainable parameters.

SRD & SRD+: The SRD dataset was created by Qu *et al*. [15]. It contains pairs of shadow and shadow-free images. In total there are 3088 pairs, where 2680 are part of the train set and 408 are part of the test set. The dataset is diverse in the following aspects: scenes, illumination conditions and casting objects with different reflectance properties and multiple silhouettes.

The dataset contains mostly images of size 840×640, so a network with bigger receptive field is required. Thus, we use all the layers of our network architecture. The number of trainable parameters for this architecture is $130k$, using 42 channels per layer. SRD images with dimension bigger than 840 were resized to have maximum dimension size of 840.

Unlike the ISTD dataset, that includes the shadow masks, there is no ground truth shadow mask for the SRD dataset. Therefore the shadow mask must be estimated. We, like previous works [4, 6], use the shadow masks calculated by DHAN [4], even though we observe that they do not capture the shadow regions accurately.

We used the same method as for the ISTD+ to create the color corrected dataset SRD+. The color correction reduces the RMSE on the non-shadow region from 12.79 for the SRD to 12.35 for the SRD+.

Evaluation metric. We use two metrics to evaluate our results - Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) similar to [11] [2]. First, the shadow-free image and predicted shadow-free image are resized to 256×256, then the error is evaluated in LAB space. The error is reported for the test set in the shadow, non-shadow and whole-image regions. Note that our network is fully convolutional and outputs the predicted shadow-free images with the same resolution as the input images. We evaluate the error metrics after resizing the images in order to be comparable to other methods.

5.2. Comparison with other methods

We compare our method against a number of competing methods. Le and Samaras [11], Liu *et al*. *Sup* [14], Fu *et al*. [6], Cun *et al*. [4] and Wang *et al*. [18] train their networks on a training data containing triplets of shadow, shadow-free and shadow masks images, similar to our method. Hu *et al*. [8] trained their network in an end-to-end manner and need only the shadow and shadow-free images for training. In [12] Le and Samaras trained their network on unpaired shadow and non-shadow patches cropped from the shadow image and thus do not need the shadow-free image. Liu *et al*. [14] synthesize pseudo shadows in the non-shadow area in the shadow image and thus need only the shadow and shadow mask for training. Gong and Cosker [7] is an interactive method that relies on the user input and does not require training.

Figure 2 shows a visual comparison on the ISTD+ dataset (we did not include works, such as [6, 18], that changed the predicted shadow free image aspect ratio). Table 1 reports the results on ISTD+ dataset [3]. As can be seen,

[2] Please note at the work of Le and Samaras [11] our MAE evaluation is referred as RMSE.

[3] Results are not exactly the same as the Authors published in their original papers since our evaluation code was written in Python, whereas others mostly used Matlab. The same python script was used for all methods. The same applies to the SRD dataset.

Shadow | Shadow Free | Ours | Le *et al.* [11] | Liu *et al.* *Sup* [14] | Le *et al.* [12] | Hu *et al.* [8] | Liu *et al.* [14]

Figure 2. **ISTD+ comparison to other methods:** (1st row) our method reconstructs the shadow area occluded by a sign very well and the non-shadow area is kept very similar to the non-shadow area of the shadow image, the other methods have reconstruction errors in the sign area. Le *et al.* [12] also add visible artifact at the non-shadow area. (2nd row) all expect ours and Hu *et al.* [8] fail to reconstruct the umbrella umbra area. We slightly outperform Hu *et al.* [8] in the penumbra area. (3rd row) our shadow area reconstruction is comparable with le *et al.* [11], and outperform other methods. (4th row) our method keeps the non-shadow area unchanged, while other methods add very visible artifacts to the black glass tiles.

Shadow | Shadow Free | Ours | Hu *et al.* [8] | Cun *et al.* [4]

Figure 3. **SRD comparison to other methods:** (1st row) The motorcycle image is a challenging scene since the plastic and seat have many black shades. As can be seen, all the methods reconstruct the shadow area in a comparable way. In the non-shadow area Cun *et al.* [4] add white artifacts under the motorcycle logo, while we preserve the original scene. (2nd row) Our method reconstructs the shadow area with the least amount of artifacts at the sand when compared to the other methods.

we outperform other methods in most metrics except MAE in the non-shadow area and RMSE in the shadow area, where we are in second place. In the non-shadow MAE we are outperformed by Gong and Cosker [7]. However, they require the user input during the inference to define the shadow and non-shadow regions. As for the shadow RMSE, we are outperformed by Fu *et al.* [6]. However, they assume that they have access to the shadow-free image when evaluating the shadow mask, whereas we are estimating this mask from the shadow image.

We repeat the experiment on the SRD dataset and report results in Table 2. We also report there our results on the SRD+ dataset, which are not that different. We come either in the first or second place in all cases. Cun *et al.* [4] that performs very well on this dataset, does not perform that well on the ISTD+ dataset. Figure 3 shows a visual com-

Method	Train Data	RMSE			MAE		
		Shadow	Non-Shadow	All	Shadow	Non-Shadow	All
Le and Samaras [11]	Shd., Shd.Free, Mask	9.68	3.33	5.77	7.19	2.91	3.61
Liu *et al.* [14] *Sup*	Shd., Shd.Free, Mask	10.08	3.55	6.16	7.43	3.03	3.75
Hu *et al.* [8]	Shd., Shd.Free	9.97	3.53	6.11	7.52	3.14	3.86
Liu *et al.* [14]	Shd., Mask	11.21	3.83	6.64	8.87	3.01	3.96
Le and Samaras [12]	Shd., Mask	13.00	3.99	6.92	9.69	2.93	4.03
Gong and Cosker [7]	No Training	16.60	4.42	7.66	12.99	**2.58**	4.28
Fu *et al.* [6]	Shd., Shd.Free, Mask	**8.78**	3.64	6.31	6.57	3.84	4.29
Cun *et al.* [4]	Shd., Shd.Free, Mask	11.28	5.38	9.31	11.33	7.18	7.86
Wang *et al.* [18]	Shd., Shd.Free, Mask	14.38	6.24	10.80	13.25	7.70	8.60
Oracle based solution	No Training (using GT images)	3.69	1.47	2.55	2.46	1.45	1.61
Ours	Shd., Shd.Free, Mask	8.91	**2.42**	**5.27**	**6.56**	2.77	**3.39**

Table 1. **Shadow removal on the ISTD+ dataset:** We report deshadow results on the ISTD+ dataset. The "Train Data" column details the data used for training by the different methods, where "Shd." means "Shadow Image" and "Shd.Free" means "Shadow Free". The best and the second best results are highlighted with **bold** font and underline, respectively. Fu *et al.* [6], that outperform us in the shadow area RMSE, compute at inference time the shadow masks using Otsu's algorithm on the difference between the shadow and shadow free images.

Method	Train Data	RMSE			MAE		
		Shadow	Non-Shadow	All	Shadow	Non-Shadow	All
Fu *et al.* [6]	Shd., Shd.Free, Mask	10.63	5.27	9.13	8.14	5.87	6.51
Hu *et al.* [8]	Shd., Shd.Free	11.67	4.40	7.62	9.14	3.56	5.14
Cun *et al.* [4]	Shd., Shd.Free, Mask	**9.41**	**3.96**	**6.85**	**7.53**	3.66	**4.75**
Ours	Shd., Shd.Free, Mask	10.53	4.03	6.99	8.23	**3.41**	4.77
Ours SRD+	Shd., Shd.Free, Mask	10.17	4.16	7.21	8.24	3.57	4.97

Table 2. **Shadow removal on the SRD dataset:** We compare our method with several other state-of-the-art methods. As can be seen, we achieve competitive results in all measures.

parison on the SRD dataset.

Our shadow removal network is quite small - only $100k$ parameters, compared to millions in competing methods. See Table 4. We use the same shadow mask detection network as previous methods did, so the overall size of both our networks is the second smallest, yet we achieve SOTA results. We only lag behind the method of Cun *et al.* [4] that is not physics based.

5.3. Ablation study

To quantify the importance of the different ways to evaluate the parameters w, b we conduct an ablation study. See Table 3. The table compares different ways to obtain w, b. As can be seen, using the ground truth mask leads to the best overall results (top row), which is to be expected. Our proposed method of predicting a binary shadow mask works slightly better than predicting a probabilistic mask, and much better than not estimating it at all. Trying to predict a shadow-free image directly works well, but not as well as our method.

In another ablation study, we investigate the gap between our solution and the optimal solution obtained by computing local w, b maps, same as in our network, but with the use of the shadow image and ground truth shadow-free image. The maps were created in the following manner. First, we calculated three coefficients maps using least-squares on a disk-shaped neighborhood sized 9, 19 and 33 pixels. Then, the maps were combined together according to weights determined by the absolute normalized cross correlation value. Values closer to one represent an area with a good linear fit and thus is weighted higher. To further improve the results around the shadow border, we follow with an aggregation step. Since each pixel belongs to several overlapping disk shaped neighborhoods we can combine the estimations of all the relevant neighborhoods by weighted average. The weight is again determined by the

Network Prediction	Mask	RMSE Shadow	RMSE Non-Shadow	RMSE All	MAE Shadow	MAE Non-Shadow	MAE All
w, b	Ground Truth	8.94	**2.17**	**4.97**	6.77	**2.56**	**3.25**
w, b	Binary	**8.91**	2.42	5.27	**6.56**	2.77	3.39
w, b	Probability	9.19	2.52	5.49	6.97	2.89	3.55
w, b	Without	12.20	3.09	7.15	9.59	3.38	4.39
Shadow Free	Binary	9.88	2.69	5.87	7.32	3.09	3.78

Table 3. **ISTD+ Ablation Study:** (1st row) using the ground truth shadow mask gives best results. Predicting a binary mask (2nd row) gives better results than predicting a probabilistic mask (3rd row). Not using a shadow mask at all (4th row) gives the worst result. Trying to predict a shadow free image directly (bottom row) leads to mediocre results.

Method	Detection (M)	Removal (M)
Le and Samaras [11]	42.5	141.2
Hu *et al*. [8]	0	79
Le and Samaras [12]	42.5	188.6
Fu *et al*. [6]	0	338.7
Cun *et al*. [4]	0	21.7
Wang *et al*. [18]	0	108.8
Ours	42.5	0.1

Table 4. **Model Size:** We and [11, 12] separate deshadowing into two smaller tasks, detection and removal. [8] learns directly from shadow to shadow free image without mask. [6] uses the ground truth shadow-free images to extract the mask which is then used at inference. [4, 18] predict the mask and shadow-free image in one network.

absolute normalized cross correlation value.

The results are presented in Table 1 as the oracle based solution. We found that there is a gap in performance between the optimal solution (2.46 MAE) and our method (6.56 MAE). Given that the previous ablation experiment (Table 3) reports little difference in the results between using the predicted shadow mask and the ground truth mask, we conclude that improving the w, b estimation is key to future progress.

5.4. Limitations

We achieve state of the art results on ISTD+ dataset, and 2nd best results on SRD, but still have some limitations. First, we observe that we have limited success in generalizing across datasets. Evaluating the SRD test set using model trained on ISTD+ train set, results in degradation in performance, compared to a model trained on SRD train set, and vice versa. Second, in some cases, our network predicts non-physical w, b values (i.e., $w{<}1$ or $b{>}0$), which might stem from errors in the data. See example in the supplemental. These maps can yield good shadow-free predictions, but may lead to poor generalization on unseen data. Finally, our method requires triplets of shadow, shadow mask and shadow-free images for training, which is difficult to collect.

6. Conclusions

We proposed a physics-based deep learning algorithm for shadow removal. Our algorithm takes as input the shadow image and a shadow mask and computes the per pixel parameters required to remove the shadow. We calculate these parameters using a local linear model and train a neural network to predict them. Remarkably, our deshadowing network is considerably smaller (about 100K compared to millions of parameters), compared to alternative methods. Empirical evidence suggests that our method produces state-of-the-art results on standard datasets.

Acknowledgements: We would like to thank Dror Alumot and Shaul Cohen from Applied Materials, Process Diagnostics and Control unit, Israel, who supported this research. We wish to acknowledge the technical assistance with the GPU environment and deep learning frameworks provided by Ilya Nelkenbaum.

References

[1] Eli Arbel and Hagit Hel-Or. Shadow removal using intensity surfaces and texture anchor points. *IEEE transactions on pattern analysis and machine intelligence*, 33(6):1202–1216, 2010. 2

[2] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. Vis. Syst*, 2(3-26):2, 1978. 1

[3] Qifeng Chen, Jia Xu, and Vladlen Koltun. Fast image processing with fully-convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2497–2506, 2017. 2, 4

[4] 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. 2, 3, 5, 6, 7, 8

[5] 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. 2

[6] 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. 2, 5, 6, 7, 8

[7] Han Gong and Darren Cosker. Interactive removal and ground truth for difficult shadow scenes. *JOSA A*, 33(9):1798–1811, 2016. 3, 5, 6, 7

[8] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2795–2808, 2019. 2, 3, 5, 6, 7, 8

[9] 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

[10] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24:109–117, 2011. 4

[11] 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, 3, 5, 6, 7, 8

[12] Hieu Le and Dimitris Samaras. From shadow segmentation to shadow removal. In *European Conference on Computer Vision*, pages 264–281. Springer, 2020. 2, 3, 5, 6, 7, 8

[13] Hieu Le, Tomas F Yago Vicente, Vu Nguyen, Minh Hoai, and Dimitris Samaras. A+ d net: Training a shadow detector with adversarial shadow attenuation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 662–678, 2018. 3

[14] 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. 2, 3, 5, 6, 7

[15] 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. 2, 3, 5

[16] 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. 2, 3

[17] 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. 4

[18] 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. 2, 3, 5, 7, 8

[19] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[20] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. 4

[21] 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. 3

[22] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–136, 2018. 2, 4