DeshadowNet: A Multi-context Embedding Deep Network for Shadow Removal

Liangqiong Qu[1,2,3] Jiandong Tian[1*] Shengfeng He[4] Yandong Tang[1] Rynson W. H. Lau[3]
[1] State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences;
[2] University of Chinese Academy of Sciences; [3] City University of Hong Kong;
[4] South China University of Technology
{quliangqiong, tianjd, ytang}@sia.cn, hesfe@scut.edu.cn, rynson.lau@cityu.edu.hk

Abstract

Shadow removal is a challenging task as it requires the detection/annotation of shadows as well as semantic understanding of the scene. In this paper, we propose an automatic and end-to-end deep neural network (DeshadowNet) to tackle these problems in a unified manner. DeshadowNet is designed with a multi-context architecture, where the output shadow matte is predicted by embedding information from three different perspectives. The first global network extracts shadow features from a global view. Two levels of features are derived from the global network and transferred to two parallel networks. While one extracts the appearance of the input image, the other one involves semantic understanding for final prediction. These two complementary networks generate multi-context features to obtain the shadow matte with fine local details. To evaluate the performance of the proposed method, we construct the first large scale benchmark with 3088 image pairs. Extensive experiments on two publicly available benchmarks and our large-scale benchmark show that the proposed method performs favorably against several state-of-the-art methods.

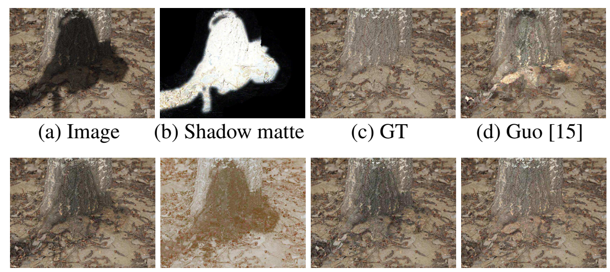
1. Introduction

The presence of illumination changes in an image, shadows in particular, have been proved to be one of the main challenging factors for a variety of computer vision tasks, such as object detection and tracking [6, 24]. As such, shadow removal aims to produce a high-quality shadow-free image given a single shadow image. According to [3, 31, 21, 1], a shadow image I_s can be considered as a pixel-wise product of a shadow-free image I_{ns} and a shadow matte S_m (or shadow scale factors).

$$I_s = S_m \cdot I_{ns}, \tag{1}$$

where the shadow matte S_m represents the illumination attenuation effects caused by the shadow.

*Corresponding author



(a) Image (b) Shadow matte (c) GT (d) Guo [15]
(e) Arbel [1] (f) Yang [33] (g) Gong [12] (h) Ours

Figure 1: Comparison with existing shadow removal methods. Existing methods fail to correctly remove the shadow cast on different semantic regions (i.e., horizontal ground and vertical trunk).

With Eq. 1, the shadow removal process is transformed to estimating a shadow matte for an input shadow image. Most existing methods follow this formulation to address the shadow removal problem [31, 21, 1, 15, 12, 19]. Notwithstanding the demonstrated success, these methods share the following three limitations.

Lack of a fully-automatic and end-to-end pipeline. Existing methods for shadow matte estimation require the prior information of shadow location. It is either obtained from shadow detection [15, 19] or user input [31, 1, 14, 12, 34]. However, shadow detection itself is a challenging task. Conventional methods for shadow detection either lack robust shadow features [35, 20, 15], or can only be applied to high-quality images [8, 30]. Due to the limited amount of training data, recent deep learning based methods [26, 19] are restricted to small network architectures.

Neglect high level semantic information. Existing works mainly adopt low-level features (e.g., color ratios [15, 12] or color statistics [31, 19]) to calculate the shadow matte. However, the shadow matte is also closely related to the semantic contents (e.g., geometry and material). As shown in Fig. 1a to 1c, the shadow matte values on two semantic regions, the horizontal ground and vertical trunk, are apparently distinct from each other since the light intensities of these two semantic regions are different. Unfortunately, exiting methods [1, 15, 12] do not consider this semantic

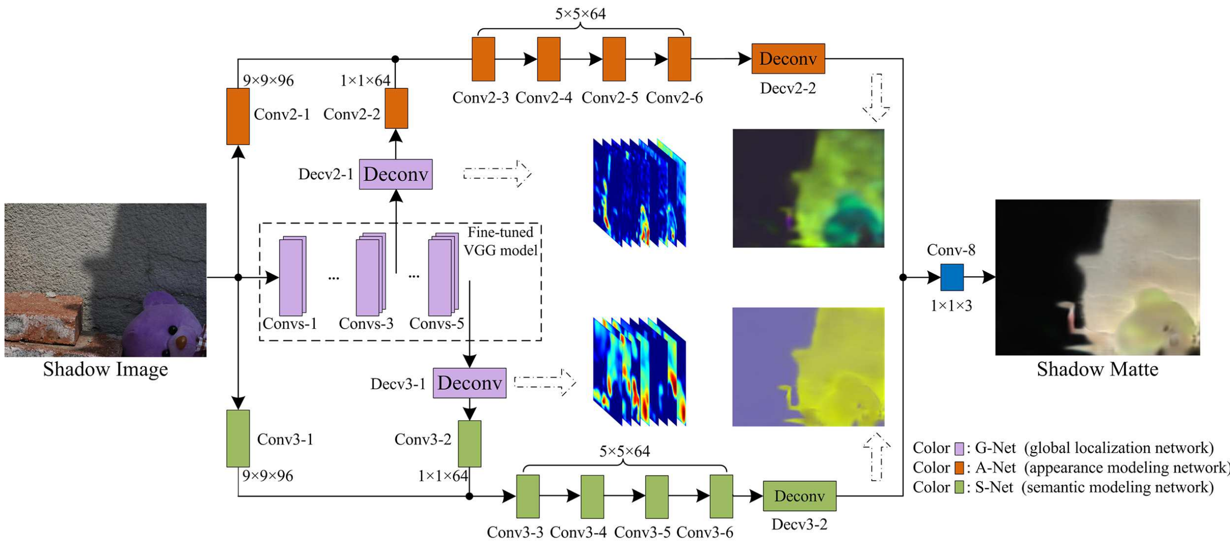


Figure 2: The architecture of DeshadowNet. DeshadowNet consists of three cooperative sub-networks: a global localization network (G-Net), an appearance modeling network (A-Net), and a semantic modeling network (S-Net). These three sub-networks are marked in different colors.

information, leading to unsatisfied results (Fig. 1d to 1g).

Require specific operation for penumbra regions. Since the content of the shadow matte may differ in umbra and penumbra regions, previous methods often adopt user-hints [5, 31] or classification (e.g., a bi-directional search [12] and thresholding operations [19]) to separate them. However, automatically identifying umbra and penumbra regions is difficult, especially for some complex background or tiny shadows (e.g., shadows of leaves).

In this paper, we aim to explore shadow removal in an end-to-end and fully automatic framework, to address the above mentioned problems. In contrast to the conventional pipeline that detect shadows, classify umbra/penumbra regions, and then remove shadows, we unify these steps into one and directly learn the mapping function between the shadow image and its shadow matte, which can then be used to recover a shadow-free image with Eq. 1. To this end, we propose a new deep neural network for shadow removal, called *DeshadowNet*. It involves a multi-context embedding mechanism, which integrates high-level semantic information, mid-level appearance information and local image details in the final prediction. The multi-context embedding is implemented by jointly training three networks, global localization network (G-Net), appearance modeling network (A-Net), and semantic modeling network (S-Net). The G-Net extracts shadow feature representation to describe the global structure and high-level semantic context of the scene. The A-Net and S-Net acquire the appearance information from the shallower layer and semantic information from the deeper layer of G-Net, respectively, allowing the prediction of fine shadow matte using multi-context information. The structure of the proposed DeshadowNet with three sub-networks is shown in Fig. 2.

To evaluate different shadow removal algorithms, we further construct a new challenging and large scale shadow removal dataset (SRD) [1]. It contains 3088 shadow and shadow-free image pairs.

2. Related work

Existing approaches for shadow removal generally include two steps: shadow localization and shadow removal. These methods first locate the shadow regions either by shadow detection [15, 19] or with user annotations [1, 14, 12, 34]. Two reconstruction algorithms with hand-crafted features are then designed for removing the detected shadows from the umbra and penumbra regions.

However, shadow detection itself is a challenging task. Conventional physically based methods can only be applied to high-quality images, while statistical learning based methods rely on hand-crafted features [12, 29, 19]. Recently, Khan *et al.* [19] and Shen *et al.* [26] take advantage of representation learning ability of Convolutional neural networks (CNNs) to learn hierarchal features for shadow detection. Due to the limited amount of training data, these two deep-based methods are restricted to small network architectures. In addition, as they apply CNNs in a patch-wise manner, a global post-processing step is required to produce consistent predictions (e.g., least-square optimization in [26] and CRF in [19]). On the contrary, DeshadowNet has a fully convolutional architecture, which can be trained end-to-end, pixels-to-pixels, to produce accurate shadow

[1] Please refer to `http://vision.sia.cn/ourteam/JiandongTian/JiandongTian.html` for the SRD dataset.

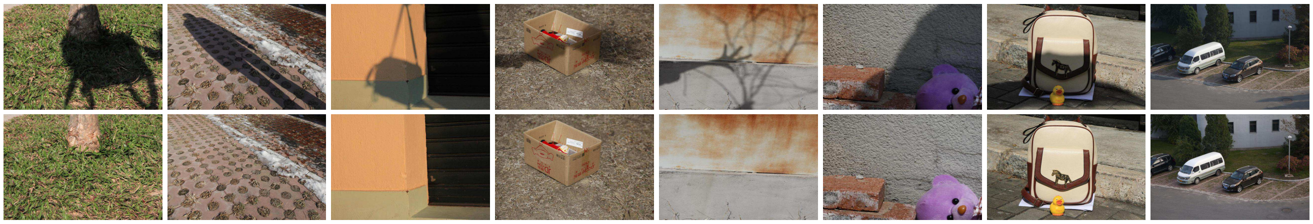


Figure 3: An illustration of several captured shadow and shadow-free image pairs in SRD.

mattes.

Even if the shadow regions are determined, the removal of shadows is still non-trivial. Existing methods remove shadows either in the gradient domain [9, 10, 23, 21] or the image intensity domain [1, 15, 12, 19]. Finlayson *et al.* [9, 10] detect shadow edges by comparing a physically deduced illumination invariant image with the original RGB image, and then propose a series of gradient domain based methods for shadow removal. These gradient based methods only modify the gradient variation in shadow edges or penumbra regions, and thus are not applicable for umbra regions with illumination variation.

On the other hand, intensity domain based shadow removal methods adopt user-hints [5, 31, 1] or classification [15, 12, 19] to determine the umbra/penumbra regions. Different low-level features are then used to estimate the shadow matte for umbra/penumbra regions respectively. Given the user annotated shadow regions, Arbel and Hel-Or [1] determine the penumbra regions using Markov Random Field, and then fit a smooth thin-plate surface model in the shadow regions to produce an approximate shadow matte. Khan *et al.* [19] first detect shadows with two separate CNNs, and then classify the umbra/penumbra regions according to gradient intensity change. Finally, they propose a Bayesian formulation to extract the shadow matte. Instead of using hand-crafted features to estimate the shadow matte, Gryka *et al.* [14] propose a Random Forest based method to model the relationship between shadow image regions and their shadow matte. Although it is a data-driven method, it requires accurate shadow annotation and an initial guess of the shadow matte as input. The final prediction is highly dependent on the initial shadow matte.

Only a few works focus on deriving a shadow-free image in an end-to-end manner. They recover the shadow-free image by intrinsic image decomposition and preclude the need of shadow detection [28, 33, 2, 25]. Strictly speaking, these intrinsic image based methods may not be considered as shadow removal methods, as they may alter the colors of the non-shadow regions (see Fig. 1f). In this paper, we propose a unified multi-context framework to embed the localization of shadows in shadow matte prediction. The proposed DeshadowNet can preserve the colors of non-shadow regions well, while removing the shadows (see Fig. 1h).

3. A New Dataset for Shadow Removal – SRD

Although the shadow removal problem has been studied for decades, publicly available datasets for this purpose are still limited. Among them, the most widely adopted shadow removal dataset is [15], which contains only 76 shadow/shadow-free image pairs. To facilitate the evaluation of shadow removal methods, we have constructed a large-scale dataset called *SRD*, which contains 3088 shadow and shadow-free image pairs. To the best of our knowledge, SRD is the first large scale benchmark for shadow removal.

To construct our dataset, we use a Canon 5D camera with a tripod and a wireless remote controller for image capturing. We set the manual capture mode with a fixed exposure parameter to capture a shadow image, where the shadow is cast by different objects. We then remove the shadow source to capture the corresponding shadow-free image. These arrangements minimize the illumination difference between two captured images.

We enrich the diversity of the proposed dataset in the following four aspects:

- **Illumination:** We take the shadow images at different illumination conditions to include hard and soft shadows in the dataset. Specifically, we capture shadows in cloudy and sunny days, and at different time of the day (e.g., dawn, morning, midday, afternoon, dusk). For example, in Fig. 3, the first two are hard shadow images and the $3rd$ to $5th$ are soft shadow images.
- **Scene:** We capture shadow images from a variety of scenes, e.g, parks, campuses, buildings, streets, mountains and beaches.
- **Reflectance:** We cast shadows on different semantic objects to obtain different reflectance phenomena. The $6th$ and $7th$ images in Fig. 3 show two examples.
- **Silhouette:** We use occluders of various shapes and geometries to cast shadows of different silhouettes and penumbra widths. The $4th$ and $5th$ images in Fig. 3 show two examples.

4. Proposed Method

The proposed multi-context embedding deep network, called DeshadowNet, is shown in Fig. 2. It aims to learn a

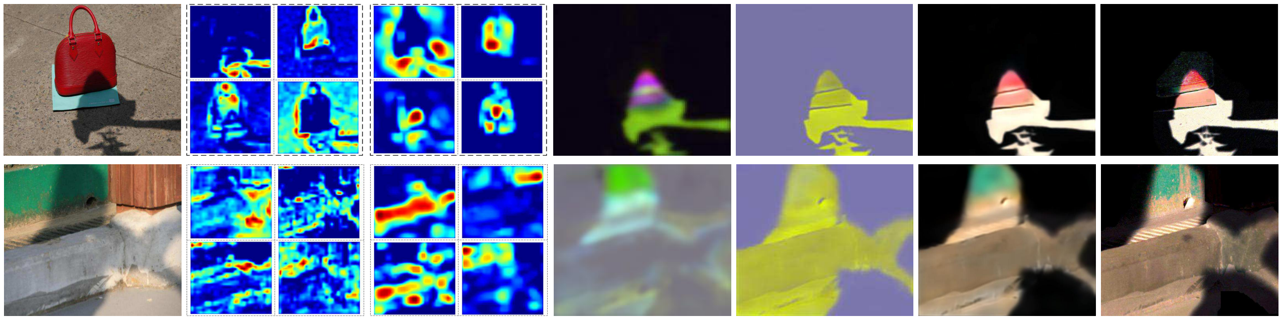


Figure 4: Visualization of the intermediate results of the proposed network. (a) shows the original shadow images and (g) shows the shadow mattes obtained from Eq. 1 and the original image pairs. (b) shows some example feature maps of the Conv3 layer of G-Net that produce (d) via A-Net, which encode appearance information of the shadow regions. (c) show some example feature maps of the Conv5 layer of G-Net that produce (e) via S-Net, which encode semantic information of shadow regions. The final predicted shadow mattes (f) embed multi-context information.

mapping function between the shadow image and its shadow matte. In this section, we first discuss the motivation and architecture of DeshadowNet, and then present the details of the training procedure.

4.1. Multi-context Convolutional Architecture

Our ideas are that an accurate shadow matte estimation method needs to understand the image content from a global perspective and model the precise illumination attenuation effects with local image details. Hence, in DeshadowNet, we implement these two ideas by designing three cooperative networks. The first network, G-Net, takes a shadow image as input and extract shadow feature representation which describes the global structure and high-level semantic information of the scene. The other two networks, A-Net and S-Net, acquire the appearance information from the shallower layer and semantic information from the deeper layer of G-Net, respectively, facilitating the prediction of fine shadow matte using multi-context information.

G-Net: Global localization network. G-Net is constructed on the basis of VGG16 network [27], which is originally designed for object recognition. Recent works suggest that CNNs trained with large amount of data on image classification task, can be well generalized across datasets and tasks such as semantic segmentation and depth prediction [11, 22, 7]. Thus, we adopt the convolutional layers of a pre-trained VGG16 model [27] and transfer their feature representation to the shadow matte prediction task with fine-tuning.

The VGG16 network contains thirteen 3×3 convolutional layers (five convolution blocks) and three fully connected layers, along with five max-pooling layers and subsampling layers. These five convolutional groups and spatial poolings substantially increase the receptive field of the network, and are thus able to extract the global context and semantic information of the scene. However, these five max-pooling layers introduce a stride of 32 pixels in the network, making the final prediction map coarse. Hence, instead of directly applying the original VGG16 architecture, we set the pooling stride to 1 in the last two max pooling layers to get a denser prediction. Except this modification, we further replace the fully-connected layers in VGG16 network with a 1×1 convolutional layers [22], followed by a deconvolution layer (see Fig. 2). These 1×1 convolutional layers enable our network to run in a fully convolutional fashion.

A-Net: Appearance modeling network / S-Net: Semantic modeling network. After extracting the global shadow features with G-Net, we then design two parallel and complementary networks (A-Net and S-Net) to predict a fine shadow matte with multi-context features.

In G-Net, each convolution block is followed by a max-polling layer, thus each of them have an progressively larger receptive field. The deeper layers of G-Net are good at capturing high-level semantic context but poor for accurate localization due to the resulted coarse features. While the shallower layers, which capture more local appearance information, cannot inject contextual information into final prediction. To better localize the shadow regions and predict fine details for the shadow matte, we further design a multi-context mechanism for local detail refinement. With this mechanism, two levels of features are derived from the G-Net and transferred to the two parallel networks (i.e., A-Net and S-Net). Specifically, while the A-Net acquires the appearance information from the shallower layer of G-Net to help model the appearance of the shadow image with local image details, the S-Net extracts the semantic information from the deeper layer of G-Net to provide semantic understanding in the final prediction. These two networks are then integrated with a convolution layer.

Fig. 4 shows some intermediate results of the proposed

Table 1: The model architecture. It takes a shadow image of resolution $8n \times 8n$ as input and outputs a shadow matte of the same size, where n is an arbitrary natural number. In this table, we set $n = 56$ (i.e., input size of 224×224) for illustration.

	Layer	convs1	convs2	convs3	convs4	convs5	Decv2-1	Decv3-1
G-Net	# of convs	2	2	3	3	3	1	1
	# of channels	64	128	256	512	512	256	256
	Filter size	3×3	3×3	3×3	3×3	3×3	8×8	8×8
	Conv. stride	1	1	1	1	1	4	4
	Zero-Padding	1	1	1	1	1	2	2
	Pool-size	2×2	2×2	2×2	2×2	2×2	-	-
	Pool-stride	2	2	2	**1**	**1**	-	-
	Output size	112×112	56×56	28×28	28×28	28×28	112×112	112×112
	Layer	conv2-1 (conv3-1)	conv2-2 (conv3-2)	conv2-3 (conv3-3)	conv2-4 (conv3-4)	conv2-5 (conv3-5)	conv2-6 (conv3-6)	Decv2-2 (Decv3-2)
A-Net (S-Net)	# of channels	96	64	64	64	64	64	3
	Filter size	9×9	1×1	5×5	5×5	5×5	5×5	4×4
	Conv. stride	1	1	1	1	1	1	2
	Zero-Padding	4	-	2	2	2	2	1
	Pool-size	3×3	-	-	-	-	-	-
	Pool-stride	2	-	-	-	-	-	-
	Output size	112×112	112×112	112×112	112×112	112×112	112×112	224×224

network. By feeding with the mid-level appearance information in Fig. 4b, A-Net predicts shadow matte in coarse scale but helps model the appearance of shadow matte (e.g., the color value of the bag and wall). On the other hand, S-Net predicts shadow matte with the guidance of high-level semantic context (e.g., semantic objects and shadows in Fig. 4c). It can predict shadow matte in fine scale (e.g., fine object boundary in Fig. 4e compared with Fig. 4d). These predicted intermediate shadow mattes demonstrate that the convolutional features in the shallower layer and deeper layer are complementary in predicting the final shadow matte. We will further analyze the effectiveness of different sub-networks in the experiment section.

To avoid the overfitting problem and achieve an optimal local minimum, drop-out is applied after each convolutional layer, and all the rectified linear Unites (ReLUs) in the networks are replaced with Parametric Rectified Linear Units (PReLUs) [17]. In contrast to ReLU, the coefficient of PReLU is adaptively learned and defined as:

$$p(x_i) = \begin{cases} x_i, & x_i \geq 0 \\ a x_i, & x_i < 0 \end{cases}, \tag{2}$$

where x_i is the input of the activation function p at channel i, and a is the learned parameter.

4.2. Training

The relationship of a shadow image I_s and its shadow matte S_m is given by Eq. 1. During training, we transform them into log space as:

$$\log(I_s) = \log(S_m) + \log(I_{ns}) \,. \tag{3}$$

Given a shadow and shadow-free image pair, we first calculate the corresponding ground-truth shadow matte S_m according to Eq. 3. Then, our goal is to learn a mapping function that infers the relationship between a shadow image and its shadow matte as:

$$S_m = F(I_s, \Theta), \tag{4}$$

where Θ represents the learned parameters of the deep network. We adopt the Mean Squared Error (MSE) as the loss function in the log space to train our model:

$$L(\Theta) = \frac{1}{K} \sum_{i=1}^{K} \left\| \log(F(I_s^i, \Theta)) - \log(S_m^i) \right\|, \tag{5}$$

where K is the total number of training samples in a batch. We minimize the loss using stochastic gradient descent (S-GD) with back-propagation.

Training strategy. Although the performance improves significantly with the increase of the network depth, training a very deep network is a non-trivial task due to the instability of the gradient vanishing/exploidng problems [4]). In this paper, we adopt the following four strategies for fast convergence and to prevent overfitting:

1. *Multi-stage training strategy.* We train DeshadowNet with two stages. The appearance and semantic streams (G-Net+A-Net and G-Net+S-Net) are first trained separately. These two streams are then connected with a convolution layer, and all three networks are jointly optimized.

2. *Multi-size training strategy.* The fully convolutional fashion of DeshadowNet enables our model to train on images of resolution $8n \times 8n$. To inject scale-invariance to the network [16], we adopt a multi-size training strategy by feeding images of three sizes:

Table 2: Quantitative results using RMSE (smaller is better). The original difference between the shadow and shadow-free images is reported in the third column. The best and second best results are marked in **red** and blue colors, respectively.

Dataset	Different regions	Original	Guo *et al.* [15]	Yang *et al.* [33]	Gong *et al.* [12]	Gryka *et al.*[14]	Khan *et al.* [19]	Ours
UIUC [15]	Shadow	42	13.9	21.6	11.8	13.9	12.1	**9.6**
	Non-shadow	**4.6**	5.4	20.3	4.9	7.6	5.1	4.8
	All	13.7	7.4	20.6	6.6	9.1	6.8	**5.9**
LRSS [14]	Shadow	44.45	31.58	23.35	22.27	-	-	**14.21**
	Non-shadow	**4.1**	4.87	19.35	4.39	-	-	4.17
	All	17.73	13.89	20.70	10.43	-	-	**7.56**
SRD	Shadow	42.38	29.89	23.43	19.58	-	-	**11.78**
	Non-shadow	**4.56**	6.47	22.26	4.92	-	-	4.84
	All	14.41	12.60	22.57	8.73	-	-	**6.64**

Table 3: Quantitative results using SSIM (larger is better). The original difference between the shadow and shadow-free image is reported in the third column. The best and second best results are marked in **red** and blue colors, respectively.

Dataset	Different regions	Original	Guo *et al.* [15]	Yang *et al.* [33]	Gong *et al.* [12]	Gryka *et al.*[14]	Ours
UIUC [15]	Shadow	0.6227	0.9228	0.8757	0.9551	0.9418	**0.9751**
	Non-shadow	**0.9861**	0.9811	0.9230	0.9839	0.9695	0.9859
	All	0.8975	0.9669	0.9114	0.9769	0.9627	**0.9832**
LRSS [14]	Shadow	0.6194	0.7905	0.8814	0.8723	-	**0.9518**
	Non-shadow	0.9882	0.9813	0.9226	0.9863	-	**0.9888**
	All	0.8637	0.9169	0.9087	0.9478	-	**0.9763**
SRD	Shadow	0.5403	0.7381	0.8601	0.8695	-	**0.9487**
	Non-shadow	**0.9843**	0.9685	0.8735	0.9790	-	0.9823
	All	0.8687	0.9087	0.8700	0.9509	-	**0.9735**

coarse scale 64×64, medium scale 128×128, and fine scale 224×224.

3. *Data synthesis.* To prevent overfitting and improve the robustness of the network, we pre-train the proposed method on a large scale synthetic shadow removal dataset. Similar to [14], we apply computer graphics techniques to synthesize shadow and shadow-free image pairs. We configure Maya with realistic light sources to project light on occluder objects, thus casting shadows on a projection plane. We have rendered 60,000 640×480 shadow/shadow-free image pairs by changing the light sources, occluder objects, and projection planes. We randomly change the shape and the position of the light source. There are 256 segmented objects in [13] are used as the occluder objects. Finally, we collect more than 1000 real images (without shadows) from the Internet as the projection plane.

4. *Data augmentation.* We augment the training data with three different operations: image translations, flipping and cropping.

Implementation. We have implemented DeshadowNet using Caffe [18]. All the networks described in this paper are trained and tested on a single NVIDIA Tesla K40m. The proposed network takes $3 \sim 5$ weeks of training to converge. The detailed configuration of DeshadowNet is shown in Table 1. The filter weights in A-Net and S-Net are initialized with random Gaussian variables (with mean value $\mu = 0$ and standard deviation $\sigma = 0.001$). We set the momentum to 0.9 and the weight decay to 0.0005 for training. The learning rate for G-Net is set to 10^{-5}. The learning rate for the rest of the network is set to 10^{-4}, and it is progressively decreased during training. In general, the proposed DeshadowNet is fast, and it takes only 0.3s to recover a shadow-free image of resolution 640×480.

5. Experiments

In this section, we extensively compare DeshadowNet with several state-of-the-art shadow removal methods on the publicly available UIUC dataset [15] and LRSS dataset [14], as well as on our proposed SRD dataset.

UIUC dataset [15]. It contains 76 shadow and shadow-free image pairs. We use all these images for testing.

LRSS dataset [14]. It contains 37 image pairs. It is specially designed to evaluate the performance on soft shadow removal. LRSS mainly contains shadows with diverse penumbra widths. We use all these images for testing.

SRD dataset. Our new dataset contains 3088 image pairs. We divide them into two parts randomly: 2680 for training and the remaining 408 images for testing.

Compared methods. We compare the proposed DeshadowNet with five state-of-the-arts methods: three automatic methods [15, 33, 19] and two interactive methods [12, 14] (requiring annotations of shadow and non-shadow

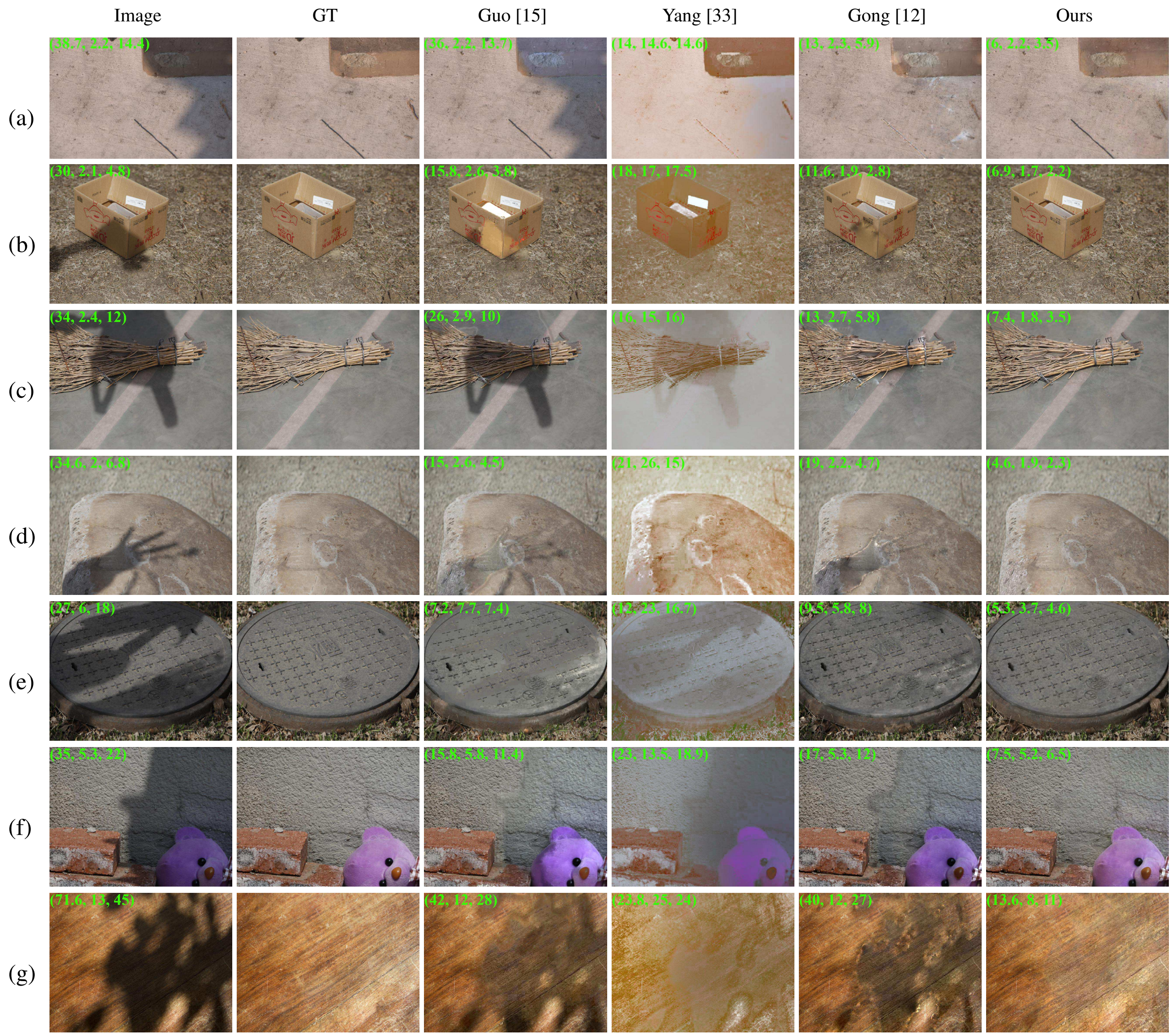


Figure 5: Shadow removal results of different methods on images with different types of shadows. The RMSE errors with respect to different regions (shadow, non-shadow, and the entire image) are marked at the top-left hand corner of each image.

regions). For fair comparison, we use either the publicly available source codes provided by the authors [15, 33, 12] or the quantitative/quanlitative results reported in the papers [19, 14] [2]. Note that [19] and [14] only report their performances on the UIUC dataset [15].

5.1. Performance Comparisons

Following the setting in [15] and [32], we adopt the root mean square error (RMSE) and structure similarity index (SSIM) in the LAB color space as evaluation metrics. While RMSE directly measures the per-pixel error between the recovered images and the ground truth images, SSIM considers structural information, which is more consistent with human visual perception.

Table 2 and 3 reports the RMSE and SSIM values, respectively, of different shadow removal methods on the UIUC dataset [15], LRSS dataset [14], and the proposed SRD test dataset. We evaluate the performance of different methods on the shadow regions, non-shadow regions, and the whole image. We can see that the proposed DeshadowNet achieves the best performance among all the compared methods and datasets.

In Fig. 5, we show some qualitatively shadow removal results from different methods, marked with their RMSE errors with respect to different regions (shadow, non-shadow, and whole image). The images in the first three rows

[2] The shadow-free results of [14] are obtained from their project website: `http://visual.cs.ucl.ac.uk/pubs/softshadows/`.

Table 4: The effectiveness of different sub-networks in DeshadowNet (measured by RMSE on the UIUC dataset [15]).

Regions	S-Net	G-Net+A-Net	G-Net+S-Net	DeshadowNet
Shadow	14.2	11.85	10.3	**9.6**
Non-shadow	5.85	5.04	4.82	**4.8**
All	7.90	6.7	6.2	**5.9**

(Fig. 5a, 5b, and 5c) contain shadows cast on different semantic regions. Both Guo [15] and Gong [12] may perform well on a specific semantic region, but fail to remove the shadows on others. For example, in Fig. 5b, Guo [15] successfully removes the shadow on the ground, but fails on the box surface. As a contrast, the proposed DeshadowNet works well on these situations by integrating multi-context information and local image details.

The images in Fig. 5d and 5e contain shadows of widely varying penumbra widths. It is difficult for existing methods [15, 12] to detect these shadows accurately. As shown in Fig. 5d, Guo [15] fails to detect the soft shadow. Even though the shadow is detected perfectly in [12] (through user annotation), the recovery of the shadow-free image is still unsatisfactory, due to the difficulty in automatically identifying umbra and penumbra regions from this image. Yang [33] can directly obtain a shadow-free image without shadow detection, but it also alters the colors of the non-shadow regions. Fig. 5f and 5g show results of different methods on more complicated situations, i.e., shadow cast on multiple semantic regions (e.g., brick, wall, and teddy bear) in Fig. 5f and the highly complex shadow in Fig. 5g.

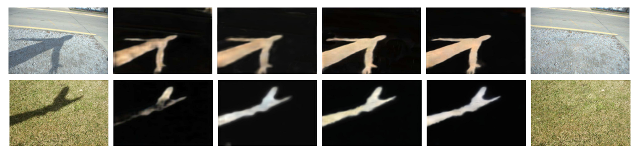
These quantitative and qualitative comparison results demonstrate that the proposed DeshadowNet can effectively recover high-quality shadow-free images from shadow images, even though with shadows cast on different semantic regions.

5.2. Component Analysis

Our DeshadowNet consists of three sub-networks, i.e., G-Net, A-Net, and S-Net. To further analyze the effectiveness and necessity of different sub-networks, we have trained three variant models of DeshadowNet and conducted a series of experiments on the UIUC shadow dataset [15]. These three variant models are: a model using S-Net only (or A-Net, since without the feature map of G-Net, A-Net is identical to S-Net in architecture), a model using G-Net and A-Net, and a model using G-Net and S-Net.

Table 4 shows the shadow removal performances of these three models on the UIUC dataset [15], measured in RMSE. We can see that none of the three models perform better than DeshadowNet. When removing G-Net, S-Net alone has relative poor performance and the RMSE error on the shadow regions reaches 14.2, compared with 10.3 by G-Net+S-Net. This demonstrates the effectiveness of the embedding mechanism in our network, where G-Net provides the appearance information and semantic context for the A-Net and S-Net, respectively.

Fig. 6 qualitatively compares these three models. Feeding with high-level semantic context and local image details, the G-Net+S-Net predicts more accurate shadow matte, as shown in Fig. 6d. On the other hand, G-Net+A-Net combines local image details with the mid-level appearance information from G-Net, predicts shadow matte in coarse scale but helps model the appearance of shadow matte (i.e., Fig. 6c obtains more accurate values in shadow matte than Fig. 6d but coarse segmentation). Thus in DeshadowNet, these multiple contextual information is incorporated for fine and accurate shadow matte prediction.



(a) Shadow (b) S-Net (c) G+A (d) G+S (e) Final (f) Result

Figure 6: The effectiveness of different sub-networks in DeshadowNet. From left to right: (a) shadow images, shadow mattes predicted by (b) S-Net, (c) G-Net+A-Net, (d) G-Net+S-Net, and (f) DeshadowNet; (g) the final shadow removal results by DeshadowNet.

6. Conclusion

In this paper, we have proposed an end-to-end DeshadowNet to recover a shadow-free image from a single shadow image. Unlike the conventional pipeline that requires shadow detection or user annotations, and removes shadows with hand-crafted features, DeshadowNet unifies these steps into one and directly learns the mapping function between the shadow image and its shadow matte. It does not require a separate shadow detection step nor any post-processing refinement step. Thus, DeshadowNet is adaptive to shadows with widely varying penumbra widths, and works well for shadows cast on different semantic regions.

The proposed multi-context embedding network, which integrates both high-level semantic context, mid-level appearance information and local image details, provides new insight for the research on low-level computer vision tasks. In the future, we will adapt and extend this multi-context embedding network to handle other complex illumination variations tasks (e.g., highlight, rain and snow removal).

Acknowledgments

This work was partially supported by the Natural Science Foundation of China under Grant Nos. 61473280, 61333019, and 91648118. The authors also thank the support by Youth Innovation Promotion Association CAS. This work was also partially supported by the Science and Technology Development Fund of Macao SAR (010/2017/A1).

References

[1] E. Arbel and H. Hel-Or. Shadow removal using intensity surfaces and texture anchor points. *IEEE TPAMI*, 33(6):1202–1216, 2011.

[2] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE TPAMI*, 37(8):1670–1687, 2015.

[3] H. Barrow and J. Tenenbaum. Recovering intrinsic scene characteristics. *Comput. Vis. Syst., A Hanson & E. Riseman (Eds.)*, pages 3–26, 1978.

[4] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[5] Y.-Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski. A bayesian approach to digital matting. In *CVPR*, volume 2, pages II–264, 2001.

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

[7] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, pages 2366–2374, 2014.

[8] G. D. Finlayson, M. S. Drew, and C. Lu. Entropy minimization for shadow removal. *IJCV*, 85(1):35–57, 2009.

[9] G. D. Finlayson, S. D. Hordley, and M. S. Drew. Removing shadows from images. In *ECCV*, pages 823–836, 2002.

[10] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew. On the removal of shadows from images. *IEEE TPAMI*, 28(1):59–68, 2006.

[11] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style trasfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016.

[12] H. Gong and D. Cosker. Interactive shadow removal and ground truth for variable scene categories. In *BMVC*. University of Bath, 2014.

[13] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007.

[14] M. Gryka, M. Terry, and G. J. Brostow. Learning to remove soft shadows. *ACM TOG*, 34(5):153, 2015.

[15] R. Guo, Q. Dai, and D. Hoiem. Paired regions for shadow detection and removal. PP(99):1–12, 2012.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, pages 346–361, 2014.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015.

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[19] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri. Automatic shadow detection and removal from a single image. *IEEE TPAMI*, 38(3):431–446, 2016.

[20] J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan. Detecting ground shadows in outdoor consumer photographs. In *ECCV*, pages 322–335. 2010.

[21] F. Liu and M. Gleicher. Texture-consistent shadow removal. In *ECCV*, pages 437–450, 2008.

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[23] A. Mohan, J. Tumblin, and P. Choudhury. Editing soft shadows in a digital photograph. *IEEE Computer Graphics and Applications*, 27(2):23–31, 2007.

[24] S. Nadimi and B. Bhanu. Physical models for moving shadow and object detection in video. *IEEE TPAMI*, 26(8):1079–1087, 2004.

[25] L. Qu, J. Tian, Z. Han, and Y. Tang. Pixel-wise orthogonal decomposition for color illumination invariant and shadow-free image. *Optics express*, 23(3):2220–2239, 2015.

[26] L. Shen, T. Wee Chua, and K. Leman. Shadow optimization from structured deep edge detection. In *CVPR*, pages 2067–2074, 2015.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. *IEEE TPAMI*, 27(9):1459–1472, 2005.

[29] J. Tian, X. Qi, L. Qu, and Y. Tang. New spectrum ratio properties and features for shadow detection. *Pattern Recognition*, 51:85–96, 2016.

[30] J. Tian and Y. Tang. Linearity of each channel pixel values from a surface in and out of shadows and its applications. In *CVPR*, pages 985–992, 2011.

[31] T.-P. Wu, C.-K. Tang, M. S. Brown, and H.-Y. Shum. Natural shadow matting. *ACM TOG*, 26(2):8, 2007.

[32] Y. Xiao, E. Tsougenis, and C.-K. Tang. Shadow removal from single rgb-d images. In *CVPR*, pages 3011–3018, 2014.

[33] Q. Yang, K. Tan, and N. Ahuja. Shadow removal using bilateral filtering. *IEEE TIP*, 21(10):4361–4368, 2012.

[34] L. Zhang, Q. Zhang, and C. Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE TIP*, 24(11):4623–4636, 2015.

[35] J. Zhu, K. G. Samuel, S. Z. Masood, and M. F. Tappen. Learning to recognize shadows in monochromatic natural images. In *CVPR*, pages 223–230, 2010.