Shadow Removal by a Lightness-Guided Network With Training on Unpaired Data

Zhihao Liu, Hui Yin, Yang Mi, Mengyang Pu, and Song Wang, *Senior Member, IEEE*

***Abstract*— Shadow removal can significantly improve the image visual quality and has many applications in computer vision. Deep learning methods based on CNNs have become the most effective approach for shadow removal by training on either paired data, where both the shadow and underlying shadow-free versions of an image are known, or unpaired data, where shadow and shadow-free training images are totally different with no correspondence. In practice, CNN training on unpaired data is more preferred given the easiness of training data collection. In this paper, we present a new Lightness-Guided Shadow Removal Network (LG-ShadowNet) for shadow removal by training on unpaired data. In this method, we first train a CNN module to compensate for the lightness and then train a second CNN module with the guidance of lightness information from the first CNN module for final shadow removal. We also introduce a loss function to further utilise the colour prior of existing data. Extensive experiments on widely used ISTD, adjusted ISTD and USR datasets demonstrate that the proposed method outperforms the state-of-the-art methods with training on unpaired data.**

***Index Terms*— Shadow removal, lightness guidance, unpaired data, GANs.**

I. Introduction

A SHADOW is a common natural phenomenon and it occurs in regions where the light is blocked. Shadow regions are usually darker with insufficient illumination and bring further complexities and difficulties to many computer vision tasks such as semantic segmentation, object detection, and object tracking [1]–[4]. These complexities and difficulties introduce more disturbances and uncertainties, which may bring more challenges when deploying some generic technologies to robotic systems [5], [6]. Although many *shadow removal* methods have been developed to recover the illumination in shadow regions, their performance is compromised given the difficulty to distinguish shadows and some darker non-shadow regions.

Compared with traditional methods [7]–[10], deep learning methods based on convolutional neural networks (CNNs) have been shown to be much more effective for shadow removal by training on annotated data. One popular approach is to use *paired data*, *i.e.*, both the shadow and shadow-free versions of an image, to train CNNs [11]–[14]. However, it is difficult and time-consuming to collect such image pairs – it usually requires a highly-controlled setting of the lighting sources, occluding objects, and cameras, as well as a strictly static scene. Data collected in such a controlled setting lacks diversity and the trained CNNs may not perform well with general images of different scenes, which may further affect the system stability in deployment.

In [15], Mask-ShadowGAN is proposed for shadow removal by training CNNs on *unpaired data*, where the shadow images and shadow-free images used for training have no correspondence, *i.e.*, they may be taken at different scenes. Its basic idea is to transform shadow removal to image-to-image translation, based on adversarial learning and cycle-consistency. Clearly, we can collect large-scale, diverse unpaired data easily, which can train CNNs with better generalisation capability for shadow removal. However, the performance of this method is still inferior to the CNNs trained on paired data, when testing on several well-known benchmark datasets. One reason lies in that Mask-ShadowGAN is directly trained over all colour channels in a single Cycle-GAN network [16], leading to a large number of parameters: this increases the difficulty of network training and optimisation.

In this paper we aim to improve the performance of Mask-ShadowGAN [15] by developing a new lightness-guided network with training on unpaired data. The basic idea of our method is to simplify the unpaired-data learning into two steps: first learn a part of simple and obvious knowledge of shadow, and then use it to guide the whole knowledge learning of shadow. In natural scenes, natural light is the main source of illumination, which plays a key role in shadow removal [7], [17] and shadow modellings [18], [19]. In most cases, shadow regions show similar chromaticity to but lower

Manuscript received June 24, 2020; revised November 30, 2020 and December 27, 2020; accepted December 28, 2020. Date of publication January 8, 2021; date of current version January 18, 2021. This work was supported in part by the Research and Development Program of Beijing Municipal Education Commission under Grant KJZD20191000402 and in part by the National Nature Science Foundation of China under Grant 51827813, Grant 61472029, Grant 61672376, and Grant U1803264. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sos S. Agaian. *(Corresponding authors: Hui Yin; Yang Mi; Song Wang.)*

Zhihao Liu and Hui Yin are with the Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China, and also with the Key Laboratory of Beijing for Railway Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: 16120394@bjtu.edu.cn; hyin@bjtu.edu.cn).

Yang Mi is with the Department of Data Science and Engineering, College of Information and Electrical Engineering, China Agriculture University, Beijing 100083, China (e-mail: miy@cau.edu.cn).

Mengyang Pu is with the Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China (e-mail: mengyangpu@bjtu.edu.cn).

Song Wang is with the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29201 USA, and also with the College of Intelligence and Computing, Tianjin University, Tianjin 300072, China (e-mail: songwang@cec.sc.edu).

Digital Object Identifier 10.1109/TIP.2020.3048677

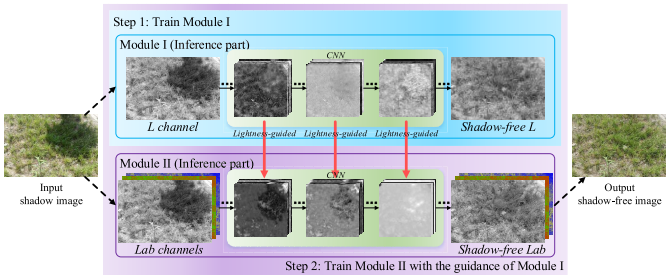


Fig. 1. An illustration of the proposed idea of training the shadow removal model with the guidance of lightness information. The first CNN module (Module I) is trained in the first step for lightness features, which are then connected to the second CNN module (Module II) in the second step for guiding the learning of shadow removal (red arrows), by further considering all colour information. Only the inference part of each module is shown.

lightness than non-shadow regions [20], [21]. Therefore, lightness is a very important cue of shadow regions. Following the above two-step idea, we propose to learn obvious but important shadow knowledge first and then use it to guide the full learning for shadow removal. This way, the difficulty of training gradually increases, as in [22], and the parameters in each step are learned separately, resulting in improved training and testing performance. We explore the lightness at feature levels instead of the input level to better represent the shadow knowledge for distinguishing the shadow regions and the dark albedo material regions that also show lower lightness and can be easily confused with shadows [23].

More specifically, by representing the input image in the *Lab* colour space [24], where *L* channel reflects the image lightness, we first train a CNN module (Module I) to compensate for the lightness in the *L* channel. As illustrated in Fig. 1, we propose to use the learned CNN features of lightness to help train a second CNN module (Module II) for shadow removal by considering all *Lab* channels. The first CNN module is connected to the second CNN module through multiplicative connections [25] to form a Lightness-Guided Shadow Removal Network (LG-ShadowNet). The multiplicative connection can combine the features of one stream with the features of the other stream, *i.e.*, combine the lightness features from Module I with the features of Module II, leading to a two-stream-like lightness-guided architecture.

Furthermore, we introduce a new loss function to further utilise the colour prior of existing data. This loss is a variant of the colour loss used for image enhancement [26], which encourages the learning of colour consistency between the generated data and the input data. Considering the performance and computational efficiency, we keep the number of parameters of LG-ShadowNet roughly the same as Mask-ShadowGAN [15] by following the strategies used for SqueezeNet [27]. In the experiments, we also discuss the use of value channel in HSV colour space for lightness feature learning given its similarity to the lightness channel in *Lab* colour space.

In short, there are two CNN modules in LG-ShadowNet and each module aims to obtain a shadow-free generator. In addition, there are three extra networks – a shadow generator, a shadow discriminator, and a shadow-free discriminator – in each module only for training the shadow-free generator adversarially [28] using cycle-consistency constraints [16]. The final shadow removal result of LG-ShadowNet is produced by the combined shadow-free generator that is formed by connecting the shadow-free generators of the two trained CNN modules.

The main contributions of this work are:

- A new lightness-guided method is proposed for shadow removal by training on unpaired data. It fully explores the important lightness information by first training a CNN module only for lightness before considering other colour information.
- An LG-ShadowNet is proposed to integrate the lightness and colour information for shadow removal through multiplicative connections. We also explore various alternatives for these connections and introduce a new loss function based on colour priors to further improve the shadow removal performance.
- Extensive experiments are conducted on widely used ISTD [14], adjusted ISTD [18] and USR [15] datasets to validate the proposed method as well as justifying its main components. Experimental results demonstrate that the proposed method outperforms the state-of-the-art methods with training on unpaired data.[1]

II. Related Work

In this section, we briefly review the related work on shadow removal, two-stream CNN networks, and image in-painting.

A. Shadow Removal

Traditional shadow removal methods use gradient [29], illumination [10], [19], [30], and region [8], [31] information to remove shadows. In recent years, deep learning methods based on CNNs have been developed for shadow removal with significantly better performance than the traditional methods. Most of them rely on paired data for supervised training. In [13], a multi-context architecture is explored to embed information from three different perspectives for shadow removal, including global localisation, appearance, and semantics. In [14], a stacked conditional generative adversarial network is developed for joint shadow detection and shadow removal. In [12], direction information is further considered to improve shadow detection and removal. In [11], an attentive recurrent generative adversarial network is proposed to detect and remove shadows by dividing the task into multiple progressive steps. In [18], a shadow image decomposition model is proposed for shadow removal, which uses two deep networks to predict unknown shadow parameters and then obtain the shadow-free image according to their decomposition model. In [32], a document image shadow removal method is developed by estimating global background colour and attention map for better recovering the shadow-free image. However, document images are quite different from natural images in terms of the variety of colours, textures and background – it might not be feasible to estimate such a consistent global background colour in natural images. To remove the

[1] All codes and results are available at https://github.com/hhqweasd/LG-ShadowNet.

reliance on paired data, a Mask-ShadowGAN framework [15] is proposed based on the cycle-consistent adversarial network of CycleGAN [16]. By introducing the generated shadow masks into CycleGAN, Mask-ShadowGAN uses unpaired data to learn the underlying mapping between the shadow and shadow-free domains. However, these CNN-based methods process all information (or all channels) of the input images together, and none of them takes out the lightness information from the input images for a separate training. In this paper, we train a CNN module exclusively for lightness before considering other colour information and the proposed LG-ShadowNet trained on unpaired data can achieve comparable performance to the state-of-the-art CNN methods trained on paired data.

B. Two-Stream Architecture

Our lightness-guided architecture are derived from two-stream architectures which have been successfully used for solving many computer vision and pattern recognition tasks [25], [33]–[40]. In [35], a two-stream processing technique is proposed to fuse the acoustic features and semantics of the conversation for emotion recognition. In [33], [38], two-stream CNN architectures are developed to combine the spatial and temporal information for video-based action recognition. In [37], a two-stream framework is proposed to combine the first-order and the second-order information of skeleton data for action recognition. In [34], a two-stream network is developed to extract garment and 3D body features, which are fused for 3D cloth draping. In [25], the motion gating is employed to the residual connections in a two-stream CNN, which can benefit action recognition. While the lightness-guided architecture of the proposed LG-ShadowNet is structurally similar to the one used in [25], they solve completely different problems: shadow removal in this paper and action recognition in [25]. Furthermore, in our LG-ShadowNet, the two modules work like a teacher (Module I) and a student (Module II) for lightness guidance and shadow removal respectively, while the two streams in [25] work simply like teammates with the same goal.

C. Image in-Painting

Our proposed method restores the illumination of the shadow region in an image, thus is also related to the long line of previous works on image in-painting. Early works [41]–[44] used hand-crafted features to fill in missing regions for image in-painting. The robustness of these works is limited on large-scale images. In recent years, deep-learning-based methods have significantly improved the performance of image in-painting. In [45], context encoders are developed to learn appearance and semantics of visual structures for image in-painting. In [46], a two-discriminator architecture is explored to enforce the visual plausibility of both the global appearance and local appearance. In [47], contextual attention is proposed to capture long-range spatial dependencies during in-painting which can find pixels from distant locations to help restore the missing regions. In [48], a foreground-aware image in-painting system is developed for better inferring the structures and completing the image content. In [49], a contextual residual aggregation mechanism is proposed to produce high-frequency residuals for missing contents. All these works need context information to restore the lost or destroyed regions, while our shadow removal method aims to restore the illumination of a shadow region where the original contents, *e.g.*, texture and edge, still exist in the shadow region.

III. METHODOLOGY

In this section, we first give an overview about our method. We then elaborate on the proposed LG-ShadowNet and loss function, as well as the network details of LG-ShadowNet. After that, we describe the details of multiplicative connections. Finally, we analyse the convergence of LG-ShadowNet.

A. Overview

The pipeline of the proposed LG-ShadowNet is shown in Fig. 2. There are two modules in LG-ShadowNet and each module aims to obtain a shadow-free generator, denoted as G_f^L and G_f^{Lab} in Module I and Module II, respectively. There are also three extra networks – a shadow generator, a shadow discriminator, and a shadow-free discriminator – in each module only for training the shadow-free generator, denoted as G_s^L, D_s^L, D_f^L and $G_s^{Lab}, D_s^{Lab}, D_f^{Lab}$ in Module I and Module II, respectively.

The inputs of LG-ShadowNet are selected from shadow image dataset and shadow-free image dataset. Each image in the datasets is represented in Lab colour space [24] and has three channels: L channel, a channel, and b channel. Module I only uses the L channel of each image as input, and the shadow and shadow-free inputs are denoted as I_s^L and I_f^L, respectively. Module II uses all Lab channels of each image as input, and the shadow and shadow-free inputs are denoted as I_s^{Lab} and I_f^{Lab}, respectively.

The training of LG-ShadowNet contains two steps. In the first step, we train Module I individually using cycle-consistency loss [16], identity loss [50], and adversarial loss [28]. In the second step, we train Module II with the guidance of Module I using the above three losses and the proposed colour loss. Both Modules I and II are involved in the second training step, and the inputs are shadow data $I_s = (I_s^L, I_s^{Lab})$ and shadow-free data $I_f = (I_f^L, I_f^{Lab})$.

In the testing stage, we only use the shadow-free generators of Modules I and II in LG-ShadowNet, as shown in the inference part of Fig. 2. The input is shadow data $I_s = (I_s^L, I_s^{Lab})$ and the output is a generated shadow-free image.

B. Proposed Network

We first train Module I as shown in the top of Fig. 2 for lightness compensation, which learns a mapping between the shadow domain and the shadow-free domain on the L channel of Lab images.

In Module I, generator G_f^L maps shadow data I_s^L to shadow-free data $\hat{I}_f^L$, which is further mapped to shadow data $\widetilde{I}_s^L$ by generator G_s^L, as illustrated in the top-left of Fig. 2. Generator G_s^L maps shadow-free data I_f^L to shadow

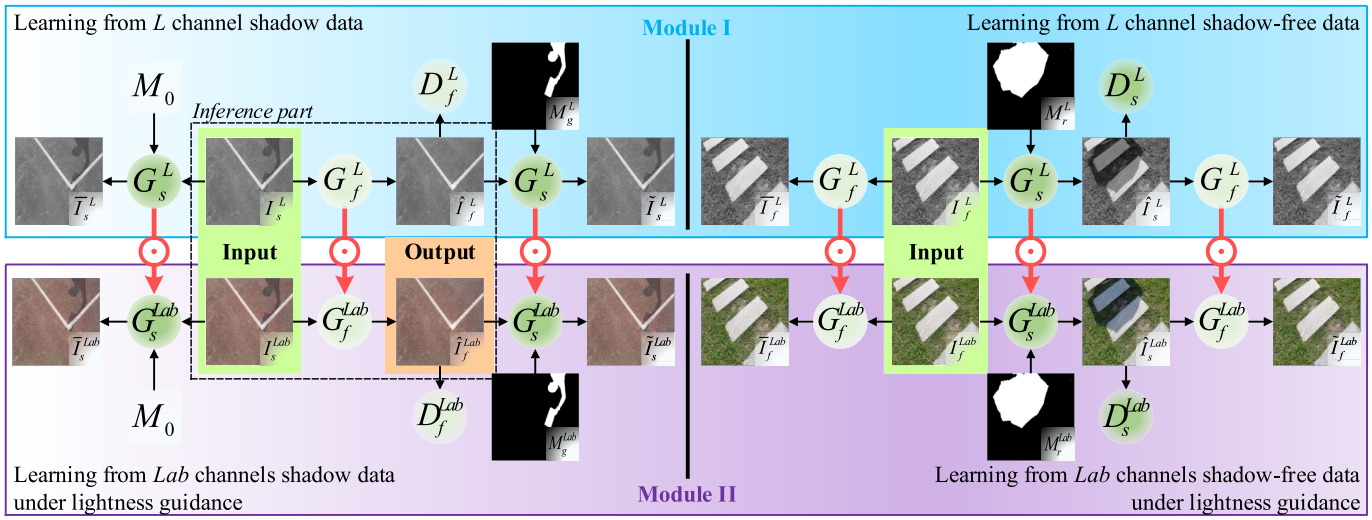


Fig. 2. An overview of the proposed LG-ShadowNet. From top to bottom show its two CNN modules, and left and right illustrate the learning from the shadow data and the shadow-free data, respectively. The generators of two modules are connected through multiplicative connections (red arrows with ⊙). We highlight the training inputs and the testing output in the green and orange boxes, respectively. The inference part in this figure is detailed in Fig. 1.

data $\hat{I}_s^L$, which is further mapped to shadow-free data $\widetilde{I}_f^L$ by generator G_f^L, as illustrated in the top-right of Fig. 2. In these processes, masks M_g^L and M_r^L are used to guide the shadow generation and are computed by following [15]. Note that M_r^L is a randomly selected mask from the previous computed M_g^L. Discriminators D_s^L and D_f^L are introduced to distinguish $\hat{I}_s^L$ and $\hat{I}_f^L$ from I_s^L and I_f^L, respectively. G_f^L also maps shadow-free data I_f^L to shadow-free data $\overline{I}_f^L$, and G_s^L also maps shadow data I_s^L to shadow data $\overline{I}_s^L$ with the guide of all-zero-element shadow-free mask M_0.

The training of Module I is the same as the training of Mask-ShadowGAN [15]. When the training is finished, we fix its parameters and move on to the learning of Module II, as shown in the bottom of Fig. 2. The parameters of generators of Module II are initialised with the parameters of Module I except for the input and output layers. Module II is connected with the Module I by multiplicative connections, resulting in the overall architecture of LG-ShadowNet. Module II takes all Lab channels as input and performs the shadow removal.

Actually, after connecting G_f^L to G_f^{Lab}, we get a new combined generator G_f. Similarly, the connection of G_s^L and G_s^{Lab} also leads to a new combined generator G_s. When learning from shadow data as shown in the left of Fig. 2, G_f converts shadow data $I_s = (I_s^L, I_s^{Lab})$ to the shadow-free data $\hat{I}_f = (\hat{I}_f^L, \hat{I}_f^{Lab})$. G_s converts the shadow-free data $\hat{I}_f$ to a generated Lab shadow image $\widetilde{I}_s^{Lab}$ with the guide of the shadow mask pair $M_g = (M_g^L, M_g^{Lab})$ that are computed from Modules I and II, respectively. This whole learning process can be summarised as

$$\widetilde{I}_s^{Lab} = G_s(G_f(I_s), M_g). \tag{1}$$

The discriminator D_f^{Lab} is introduced to distinguish $\hat{I}_f^{Lab}$ from I_f^{Lab}. G_s maps I_s to $\overline{I}_s^{Lab}$ with the guide of shadow-free mask M_0.

When learning from shadow-free data as shown in the right of Fig. 2, one of the inputs is shadow-free data $I_f = (I_f^L, I_f^{Lab})$. G_s converts I_f to a shadow data $\hat{I}_s = (\hat{I}_s^L, \hat{I}_s^{Lab})$ with the guide of the other input: mask pair $M_r = (M_r^L, M_r^{Lab})$, which are randomly selected from the previous obtained M_g. G_f converts $\hat{I}_s$ to a generated Lab shadow-free image $\widetilde{I}_f^{Lab}$:

$$\widetilde{I}_f^{Lab} = G_f(G_s(I_f, M_r)). \tag{2}$$

Discriminator D_s^{Lab} is utilised to distinguish $\hat{I}_s^{Lab}$ from I_s^{Lab} and G_f maps I_f to $\overline{I}_s$.

In short, the training of LG-ShadowNet can be briefly described as two steps: first train Module I and then train Module II with the guidance of Module I. Note that the Module I and Module II are individually trained. Previous works [51], [52] have proved that using L channel to adjust the lightness of the shadow region is effective, although it cannot adjust all colours of the shadow region perfectly. While in our method, Module I provides the lightness information to guide the learning of Module II, and only the latter performs shadow removal. The effectiveness of the lightness information is verified in our ablation experiments. In addition, if Module I and Module II are jointly trained, the performance of LG-ShadowNet slightly drops and we discuss these results in our experiments.

C. Loss Function

Following [15], we combined four losses: identity loss $L_{identity}$ [50], cycle-consistency loss L_{cycle} [16], adversarial loss L_{GAN} [28], and colour loss for training the proposed network, *i.e.*,

$$\begin{aligned}\mathcal{L}(G_f, G_s, D_f^{Lab}, D_s^{Lab}) \\ = \omega_1 L_{identity} + \omega_2 L_{cycle} + \omega_3 L_{GAN} + \omega_4 L_{colour}.\end{aligned} \tag{3}$$

The weights are experimental hyper-parameters and control the relative importance of each loss. Since we use Mask-ShadowGAN [15] as the baseline model to build our own

model, we simply follow its weight setup by setting the first three weights ω_1, ω_2, and ω_3 to be 5, 10, and 1, respectively. We empirically set ω_4 to be 10 based on two considerations. First, the weight ω_4 of the colour loss plays a similar role as the weight ω_2 of the cycle loss because both of them encourage $\widetilde{I}_s^{Lab}$ and $\widetilde{I}_f^{Lab}$ to be the same as I_s^{Lab} and I_f^{Lab}, respectively. Second, as detailed in the later experiments, we try different values of ω_4 to train the proposed LG-ShadowNet and $\omega_4 = 10$ leads to the best performance. The generators and discriminators are obtained by solving the mini-max game

$$arg \min_{G_f, G_s} \max_{D_f^{Lab}, D_s^{Lab}} \mathcal{L}(G_f, G_s, D_f^{Lab}, D_s^{Lab}). \quad (4)$$

We define the four losses in Eq. (3) as follows.

Identity loss encourages $\overline{I}_s^{Lab}$ and $\overline{I}_f^{Lab}$ to be the same as I_s^{Lab} and I_f^{Lab}, respectively:

$$\begin{aligned} L_{identity}(G_s, G_f) &= L^s_{identity}(G_s) + L^f_{identity}(G_f) \\ &= \mathbb{E}_{I_s \sim p(I_s)}\left[\left\| G_s(I_s, M_0), I_s^{Lab} \right\|_1 \right] \\ &\quad + \mathbb{E}_{I_f \sim p(I_f)}\left[\left\| G_f(I_f), I_f^{Lab} \right\|_1 \right], \end{aligned} \quad (5)$$

where $\|.\|_1$ represents the L_1 loss, p denotes the data distribution, $I_s \sim p(I_s)$ indicates I_s is selected from the data distribution p over the shadow image dataset, and $I_f \sim p(I_f)$ indicates I_f is selected from the data distribution p over the shadow-free image dataset. In theory, the data distribution is an empirical distribution [16], [28].

Cycle-consistency loss encourages $\widetilde{I}_s^{Lab}$ and $\widetilde{I}_f^{Lab}$ to be the same as I_s^{Lab} and I_f^{Lab}, respectively:

$$\begin{aligned} L_{cycle} &= L^s_{cycle}(G_f, G_s) + L^f_{cycle}(G_s, G_f) \\ &= \mathbb{E}_{I_s \sim p(I_s)}\left[\left\| G_s(G_f(I_s), M_g), I_s^{Lab} \right\|_1 \right] \\ &\quad + \mathbb{E}_{I_f \sim p(I_f)}\left[\left\| G_f(G_s(I_f, M_r)), I_f^{Lab} \right\|_1 \right]. \end{aligned} \quad (6)$$

Adversarial loss matches the data distribution over real Lab images and the data distribution over the generated Lab images:

$$\begin{aligned} &L_{GAN}(G_s, G_f, D_s^{Lab}, D_f^{Lab}) \\ &= L^s_{GAN}(G_s, D_s^{Lab}) + L^f_{GAN}(G_f, D_f^{Lab}) \\ &= \mathbb{E}_{I_s^{Lab} \sim p(I_s)}[\log(D_s^{Lab}(I_s^{Lab}))] \\ &\quad + \mathbb{E}_{I_f \sim p(I_f)}[\log(1 - D_s^{Lab}(G_s(I_f, M_r)))] \\ &\quad + \mathbb{E}_{I_f^{Lab} \sim p(I_f)}[\log(D_f^{Lab}(I_f^{Lab}))] \\ &\quad + \mathbb{E}_{I_s \sim p(I_s)}[\log(1 - D_f^{Lab}(G_f(I_s)))]. \end{aligned} \quad (7)$$

Colour loss encourages the colour in $\widetilde{I}_s^{Lab}$ and $\widetilde{I}_f^{Lab}$ to be the same as I_s^{Lab} and I_f^{Lab}, respectively:

$$\begin{aligned} L_{colour} &= L^s_{colour}(G_f, G_s) + L^f_{colour}(G_s, G_f) \\ &= \sum_p (J - \cos < (G_s(G_f(I_s), M_g))_p, (I_s^{Lab})_p >) \\ &\quad + \sum_p (J - \cos < (G_f(G_s(I_f, M_r)))_p, (I_f^{Lab})_p >), \end{aligned} \quad (8)$$

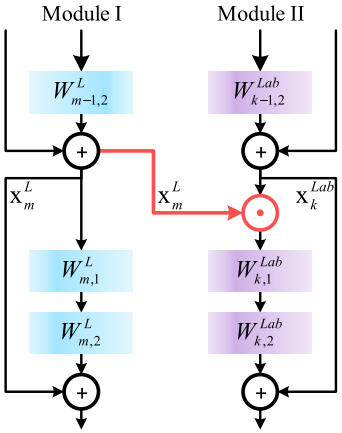


Fig. 3. An illustration of the multiplicative connections between the two modules.

where $()_p$ represents a pixel, J denotes an all-ones matrix with the same size as the input image, and cos $<,>$ represents the angle cosine between vectors. Each pixel of the generated image or input image is regarded as a 3D vector that represents the Lab colour. The angle cosine between two colour vectors equals to 1 when the vectors have the same direction. The only difference between our loss and the colour loss formulated in [26] is the latter calculates the angle rather than the angle cosine of each pixel, in which the angle between two colour vectors equals to 0 when the vectors have the same direction. We choose this colour loss by following [26], which shows that the use of angle cosines in the loss can effectively enhance under-exposed images.

D. Network Details

The architectures of Module I and II are based on Mask-ShadowGAN [15], which has two generators and two discriminators. Each generator contains three convolutional layers for input and down-sampling operations, followed by nine residual blocks with the stride-two convolutions and another three convolutional layers for up-sampling and output operations. The residual blocks [53] are derived from [54] following the architecture of [55], which has been successfully used for style transfer and super-resolution tasks. Discriminators are based on PatchGAN [56]. Instance normalisation [57] is used after each convolution layer. While the general structure of the backbone network is the same as Mask-ShadowGAN, the number of parameters is different. The original architecture of Mask-ShadowGAN is drawn from CycleGAN [16], which is designed for general image-to-image translation instead of specifically for shadow removal. In our backbone network, we follow the principle of SqueezeNet [27] to reduce the channels of Mask-ShadowGAN by half to consider both performance and efficiency.

E. Multiplicative Connections

Figure 3 shows the details of multiplicative connections used for lightness guidance between the residual blocks of two modules, in which the feature maps from Module I are

connected with the feature maps of Module II by element-wise multiplication and then sent to the weight layer of next residual block. The multiplicative connections do not affect the identity mapping x_k^{Lab} and its calculation can be written as:

$$x_{k+1}^{Lab} = x_k^{Lab} + \mathcal{F}(x_k^{Lab} \odot x_m^L, W_k^{Lab}), \tag{9}$$

where x_k^{Lab} is the input of the k-th layer of Module II, the function $\mathcal{F}$ represents the residual mapping to be learned, $\odot$ represents the element-wise multiplication, x_m^L is the input of the m-th layer of Module I and also the input of the k-th layer of Module II, and W_k^{Lab} denotes the weights of the k-th layer residual unit in Module II.

F. Convergence Analysis

The training of LG-ShadowNet consists of two steps, and we start from the first step, *i.e.*, the training of Module I. In Module I, if the generator G_f^L and the discriminator D_f^L have enough capacity, the distribution over generated shadow-free data $p(\widetilde{I}_f^L)$ shall converge to the distribution over real shadow-free data $p(I_f^L)$ according to the convergence of GANs [28].

In the second step, G_f^L provides the lightness information to Module II, which only brings signal changes to the feature maps of Module II (as shown in Eq. (9)). G_f in LG-ShadowNet is still trained to match the distribution over the real shadow-free data. Assuming we obtain the optimal generator G_f and discriminator D_f in LG-ShadowNet, since $\widetilde{I}_f^{Lab} \sim G_f(\widetilde{I}_f^{Lab}|I_f)$, when I_f is applied to G_f, we get $\widetilde{I}_f^{Lab}$ which has the same distribution as the real shadow-free data I_f^{Lab}, and the distribution over the generated shadow-free data $p(\widetilde{I}_f^{Lab})$ converges to the distribution over the real shadow-free data $p(I_f^{Lab})$.

Likewise, the distribution of the data generated by G_s converges to the distribution of real shadow data $p(I_s^{Lab})$, and the distributions of data generated by LG-ShadowNet converge to the distributions of the real shadow and shadow-free data, respectively. In practice, if Module I provides appropriate lightness information in LG-ShadowNet, the model will converge better; on the contrary, if the lightness information provided by Module I does not guide well the learning of shadow removal, the model will converge worse.

IV. Experiments

A. Datasets and Metrics

In this section, we validate our approach on three widely used shadow removal datasets:

1) ISTD [14]: It contains 1,870 image triplets with 1,330 triplets for training and 540 for testing, where a triplet consists of a shadow image, a shadow mask, and a shadow-free image. ISTD shows good variety in terms of illumination, shape, and scene;

2) Adjusted ISTD (AISTD, [14], [18]): In [18], original shadow-free images in ISTD are transformed to colour-adjusted shadow-free images via a linear regression method [18] to mitigate the colour inconsistency between the shadow and shadow-free image pairs. In this adjusted dataset, the illumination noises in the original shadow-free images are significantly reduced, *e.g.*, the Root-Mean-Square Error (RMSE) for the whole testing set of ISTD is reduced from 6.8 to 2.6 [18], and the methods trained on the adjusted ISTD dataset can also perform better with much lower RMSE. We regard this dataset as a new dataset in the following experiments. AISTD has the same shadow images, shadow masks, and training/testing data splits as ISTD;

3) USR [15]: It contains 2,445 shadow images with 1,956 images for training and 489 for testing. It also contains 1,770 shadow-free images for training. This is an unpaired dataset that covers a thousand different scenes with great diversity. There is no corresponding shadow-free image for the shadow images.

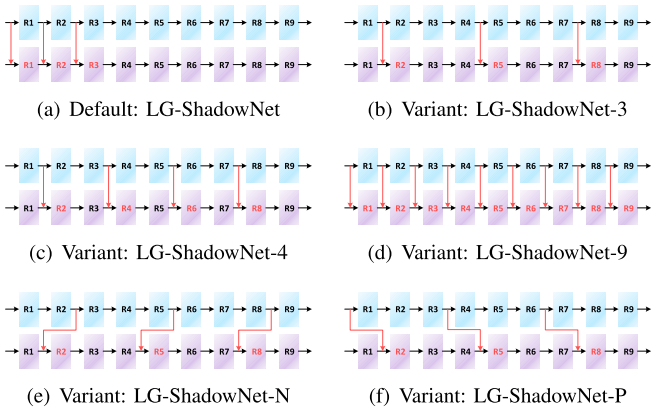
Evaluation Metrics: On the ISTD and AISTD datasets, we use the Root-Mean-Square Error (RMSE), Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) as the evaluation metrics. Following recent works [12]–[15], [18], we compute the RMSE between the ground truth images and generated shadow removal results in *Lab* colour space, at the original scale 480×640. Note that we compute RMSE *on each image* and then average the score over all images on shadow and non-shadow regions for emphasising more the quality of each image. This is more consistent with other metrics such as PSNR and SSIM.

On the unpaired dataset USR, we use several blind image quality assessment metrics and conduct the user study to evaluate the visual quality of shadow removal results, because this dataset has no ground truth for computing RMSE [15]. Specifically, we adopt the broadly used SSEQ [58], NIQE [59], and DBCNN [60] as blind image quality assessment metrics and compute each metric on each shadow removal result and then average the scores over all the results. For the user study, we recruited five participants with average age of 26. When comparing two methods, we randomly select 30 test images for each participant. For each image, he/she compares the shadow-removal results from the two methods and then votes for the better one. We then count the proportion of the 150 votes that are received by each of the two methods as their relative performance: the higher the proportion, the better its shadow-removal quality. The following experimental results achieved by our method on different datasets are trained on corresponding datasets respectively.

B. Implementation Details

Our model is initialised following a zero-mean Gaussian distribution with a standard deviation of 0.02. The model is trained by using Adam optimiser [61] and a mini-batch size is set to 1. Each sample in the training dataset is resized to 448×448 and a random crop of 400×400 is used for training which prevents the model from learning spatial priors that potentially exist in the dataset [62].

For Module I of our method, we empirically set the training epochs to 200, 200, and 100 for ISTD, AISTD, and USR, respectively. For Module II and the variants to be discussed in later experiments, we empirically set the training epochs to 100 on all datasets. Since we use Mask-ShadowGAN [15] as



(a) Default: LG-ShadowNet (b) Variant: LG-ShadowNet-3

(c) Variant: LG-ShadowNet-4 (d) Variant: LG-ShadowNet-9

(e) Variant: LG-ShadowNet-N (f) Variant: LG-ShadowNet-P

Fig. 4. Variants of the multiplicative connections between two modules. We show the nine residual blocks (R1-R9) of the generator where the convolutional layers are not shown for simplicity. The multiplicative connections are highlighted in red arrows.

the baseline model to build our own model, we simply follow its learning rate setup for all the models in our experiments. The basic learning rate is set as 2×10^{-4} for the first half of epochs and is reduced to zero with a linear decay in the next half of epochs.

Finally, our method is implemented in PyTorch on a computer with an Intel Xeon E5-2683 CPU and a single NVIDIA GeForce GTX 1080 GPU and is evaluated in MATLAB R2016a. The training time depends on the number of training epochs and the total number of training images in the training set. It takes about 67.4, 67.4, and 68.2 hours to train LG-ShadowNet on ISTD, AISTD, and USR, respectively. The testing time depends on the number of testing images in the testing set. It only takes about 5.6, 5.6, and 3.5 minutes to test LG-ShadowNet on ISTD, AISTD, and USR, respectively.

C. Variants of Two-Module Connections

To study the impact of using multiplicative connections, we try several variants of connections. These variants are shown in Fig. 4. The default connections in LG-ShadowNet, as shown in Fig. 4(a), are inserted at the first three shallow residual blocks. No connections are inserted in middle or deeper blocks since they have been shown to hurt the performance: deeper layers are specialised for one task and may not provide useful information to another task [63]. In our method, Module I is trained to compensate for the lightness in the given L channel of the shadow image, while Module II is trained to remove shadows and restore all colour information in the given shadow image. The latter performs a more difficult task than the former. Our later ablation study also shows that the variants by inserting more connections in deeper layers lack robustness.

The following three variants shown in Fig. 4(b)-4(d) connect corresponding layers in both modules, which indicates the case that $k = m$ in Eq. (9). These three variants show different intervals and different numbers of connections. Specifically, LG-ShadowNet-3 inserts connections after the first of every three residual blocks, *i.e.*, with an interval of three, similar to the setting in the Spatiotemporal Multiplier Networks [25]. Similarly, LG-ShadowNet-4 and LG-ShadowNet-9 insert the connections between the residual blocks with an interval of two and one, respectively. The two variants of LG-ShadowNet-N and LG-ShadowNet-P, shown in Fig. 4(e)-4(f), connect between non-corresponding residual layers, *i.e.*, $k \neq m$ in Eq. (9). They insert the connections from the next ($(k+1)$-th) and previous ($(k-1)$-th) residual blocks of Module I to the current (k-th) residual blocks of Module II, respectively.

We also study the impact of additive connections, *i.e.*, using addition instead of the element-wise multiplication with $k = m$ in Eq. (9) in LG-ShadowNet. This is an alternative connection used in the action recognition task [25] and may perform better than the multiplicative connection.

D. Ablation Study

We first perform an ablation study on AISTD to evaluate the effectiveness of the proposed lightness-guided architecture trained with unpaired data in different colour spaces. We use the default connection in LG-ShadowNet and remove the proposed colour loss to train different models with different inputs with different number of input channels. Note that Module II is trained by following the settings of LG-ShadowNet without the guidance of Module I and the colour loss. Besides, to verify the effectiveness of the lightness-guided architecture trained on paired data, we use a generator that maps shadow-data to shadow-free data and the L_1 loss to train Modules I and II on paired data in a *fully supervised* manner, and these models are denoted with a suffix *Sup.*. Quantitative results in terms of RMSE metric are shown in Table I and the RMSE between the real shadow and shadow-free pairs on AISTD is shown in the first row of the table.

From the second and third rows of Table I, we can see that Module I trained on V channel and L channel can significantly reduce the RMSE of the original data (row 1). The latter demonstrates the effectiveness of using L channel for lightness compensation. The results in rows 4-6 show that, using data on Lab colour space as training data is more suitable for shadow removal than using RGB or HSV colour spaces. The results in rows 7-9 show that, using L channel as the input of Module I achieves the best results than using the V channel and the Lab data. This confirms that the benefits are from the learned lightness information and the superiority of using L channel data to guide the learning of shadow removal. Compared with Module II* trained on Lab data (row 6), LG-ShadowNet* trained on $L + Lab$ data (row 9) can reduce RMSE by 8.3% from 5.64 to 5.17, which proves the effectiveness of the lightness-guided architecture. In addition, the results in rows 10-12 show that training above modules on paired data in a fully supervised manner can further improve the performance. The advantage of using L channel as the guidance and the effectiveness of the lightness-guided architecture are further verified here.

Figure 5 shows some visual comparison results of Module I trained on L data (Module I-L), Module II, and LG-ShadowNet* on the ISTD, AISTD and USR datasets. From the second and third columns, we can see that Module I-L can restore the shadow regions on L channel

TABLE I

QUANTITATIVE RESULTS ON AISTD IN TERMS OF RMSE. *Lab*, HSV AND RGB INDICATE THE RESPECTIVE COLOUR SPACES. S AND N REPRESENT THE RMSE OF SHADOW REGION AND NON-SHADOW REGION, RESPECTIVELY. '⋆' DENOTES TRAINING WITHOUT COLOUR LOSS, WHICH WE USE IN ALL THE REMAINING EXPERIMENTS. L AND V REPRESENT THE L CHANNEL OF *Lab* AND THE V CHANNEL OF HSV, RESPECTIVELY. MODULE I IS TRAINED TO COMPENSATE FOR THE LIGHTNESS ON L CHANNEL SO WE ONLY SHOW THE RMSE OF L CHANNEL FOR EVALUATING THIS MODULE

Models	Training data	*Lab*	N	S	L	N	S
Original data	-	9.16	3.33	38.53	6.05	1.61	28.09
Module I	V	-	-	-	3.70	2.26	10.73
	L	-	-	-	3.15	2.37	7.06
Module II⋆	RGB	5.78	4.74	12.03	3.23	2.49	7.48
	HSV	7.71	6.37	15.84	3.57	2.77	7.96
	Lab	5.64	4.66	11.65	3.35	2.71	7.24
LG-ShadowNet⋆	$Lab + Lab$	5.40	4.33	12.04	3.11	2.39	7.42
	V+*Lab*	5.29	4.25	11.27	3.04	2.31	7.05
	$L + Lab$	**5.17**	**4.09**	**11.15**	**2.98**	**2.23**	**6.93**
Module I *Sup.*	L	-	-	-	2.98	2.28	6.71
Module II⋆ *Sup.*	*Lab*	5.08	4.07	11.28	2.88	2.22	6.83
LG-ShadowNet⋆ *Sup.*	$L + Lab$	**4.83**	**3.95**	**10.07**	**2.66**	**2.10**	**5.83**

Fig. 5. Visual comparisons on ISTD, AISTD and USR. Three rows from top to bottom show results for one sample from ISTD, AISTD and USR, respectively. The first two columns show the input shadow images and their L channels, respectively.

TABLE II

QUANTITATIVE RESULTS ON ISTD, AISTD AND USR. EACH RESULT ON USR REPRESENTS THE PROPORTION OF VOTES RECEIVED BY THE PROPOSED LG-ShadowNet⋆ OR ITS VARIANTS WHEN COMPARED WITH MODULE II⋆ TRAINED ON *Lab* OR RGB DATA. THE SUFFIX '-*Lab*' AND '-RGB' IN THE MODEL NAME REPRESENT THE TRAINING DATA ON *Lab* AND RGB COLOUR SPACES, RESPECTIVELY, WHICH WE USE IN ALL THE REMAINING EXPERIMENTS

Models	ISTD			AISTD			USR	
	Lab	N	S	*Lab*	N	S	Module II⋆-*Lab*	Module II⋆-RGB
LG-ShadowNet⋆	6.64	6.00	10.98	**5.17**	**4.09**	11.15	64.7%	88.7%
LG-ShadowNet-3⋆	6.80	6.19	**10.70**	5.34	4.26	**11.11**	61.3%	91.3%
LG-ShadowNet-4⋆	6.69	6.06	10.78	5.37	4.26	11.17	61.3%	87.3%
LG-ShadowNet-9⋆	6.91	6.26	11.34	5.46	4.28	11.34	**72.0%**	**92.7%**
LG-ShadowNet-N⋆	**6.61**	**5.94**	10.93	5.39	4.29	11.18	64.0%	90.7%
LG-ShadowNet-P⋆	6.80	6.14	11.30	5.39	4.28	11.37	54.7%	88.7%
LG-ShadowNet-A⋆	6.89	6.13	11.58	5.26	**4.09**	11.57	52.0%	79.3%

effectively. LG-ShadowNet⋆ can produce better results than individual modules, *e.g.*, it successfully removes the shadow and restores the lightness on the top right of the image shown in the second row.

Next, we perform another ablation study on ISTD, AISTD and USR to evaluate the various connection variants of LG-ShadowNet described in subsection IV-C. All models are trained on unpaired $L + Lab$ data without the proposed colour loss. Quantitative results in terms of RMSE metric on ISTD and AISTD and the user study results of LG-ShadowNet and its variants against Module II on USR are shown in Table II.

From Tables II, we can see that different variants achieve different results on different datasets. We observe that LG-ShadowNet-N⋆ achieves better performance on the ISTD

TABLE III

QUANTITATIVE RESULTS OF VARIANTS BY INSERTING CONNECTIONS IN SHALLOWER OR DEEPER LAYERS ON AISTD IN TERMS OF RMSE

Models	Lab	N	S
LG-ShadowNet-9*	5.46	4.28	11.34
LG-ShadowNet-Deep*	5.34	4.15	11.45
LG-ShadowNet*	**5.17**	**4.09**	**11.15**

dataset. This indicates that inserting multiplicative connections between deeper layers of Module I and shallower layers of Module II could be more effective.

On AISTD, LG-ShadowNet* significantly surpasses other variants, which shows that low-level features are sufficient for guiding the learning of shadow removal, while embedding more high-level features leads to inferior results, especially in non-shadow regions. The variant of LG-ShadowNet-9* achieves the best result on USR, which means using the lightness features from deeper layers may be more effective in producing visually pleasing results. However, from the results of LG-ShadowNet-9* on ISTD and AISTD datasets, we can see that inserting connections in middle and deeper blocks leads to worse results, indicating its lack of robustness. In addition, we try a variant LG-ShadowNe-Deep* that inserts connections in deeper blocks (the last three blocks) and trained it on the AISTD dataset. The results are also shown in Table III. We observe that this variant fails to surpass the best variant, which further proves that inserting connections in deeper blocks performs worse than inserting them in shallower blocks.

Comparing LG-ShadowNet* with the LG-ShadowNet-A*, we observe that the additive connections lead to inferior performance. Also, when we compare LG-ShadowNet-A* with Module II, we observe that LG-ShadowNet-A* achieves lower RMSE values. The above comparisons show that, although LG-ShadowNet-A* performs worse than LG-ShadowNet*, it outperforms the Module II, which means that the variant using additive connections is better than the one without using any connection, but its performance is not as good as the variant using multiplicative connections. One possible reason might be that multiplicative connections can bring stronger or better guidance signals than additive connections as pointed out in [25]. We chose LG-ShadowNet as the default connection variant because it performs more robust than other variants on all three datasets.

We report the qualitative results of LG-ShadowNet trained with and without the proposed colour loss in Table IV to evaluate the effectiveness of the proposed colour loss. On the ISTD dataset, we can see that the colour loss has little effect on the overall RMSE, but it improves the RMSE of non-shadow regions, *i.e.*, the quality of most parts of the results is improved. Comparing the statistics on AISTD and USR, we observe the conspicuous improvement by using the proposed colour loss. We also try different values for the weight ω_4 of the colour loss in Eq. (3) to train the proposed LG-ShadowNet and find that $\omega_4 = 10$ leads to the best performance, as shown in Table V. On the whole, the colour loss that restricts the colour direction to be the same is an effective constraint for shadow removal.

TABLE IV

QUANTITATIVE RESULTS OF LG-SHADOWNET TRAINED WITH AND WITHOUT THE COLOUR LOSS ON ISTD, AISTD AND USR

Models	ISTD			AISTD			USR
	Lab	N	S	Lab	N	S	Lab
Ours w/o L_{colour}	**6.64**	6.00	**10.98**	5.17	4.09	11.15	36.7%
Ours with L_{colour}	6.67	**5.91**	11.63	**5.02**	**4.02**	**10.64**	**63.3%**

TABLE V

QUANTITATIVE RESULTS OF THE PROPOSED LG-SHADOWNET BY USING DIFFERENT VALUES OF ω_4 ON AISTD IN TERMS OF RMSE

ω_4	0.1	1	10	100
RMSE	5.42	5.20	**5.02**	6.15

TABLE VI

QUANTITATIVE RESULTS OF THE PROPOSED LG-SHADOWNET BY USING DIFFERENT BASE LEARNING RATES ON AISTD IN TERMS OF RMSE

Base Learning Rates	2×10^{-5}	2×10^{-4}	2×10^{-3}
RMSE	5.58	**5.02**	9.06

TABLE VII

QUANTITATIVE RESULTS OF THE PROPOSED LG-SHADOWNET AND THE COMPARED MASK-SHADOWGAN IN TERMS OF SSEQ, NIQE, AND DBCNN METRICS ON THE USR DATASET. FOR THESE THREE METRICS, THE LOWER THE BETTER

Methods	NIQE[59]	SSEQ[58]	DBCNN[60]
Mask-ShadowGAN-RGB [15]	5.06	28.67	40.10
Mask-ShadowGAN-Lab [15]	4.87	28.22	39.90
LG-ShadowNet (Ours)	**4.70**	**27.49**	**39.58**

To justify the choice of the base learning rate, we train our model using different base learning rates on the AISTD dataset. The results in Table VI show that overly large or small learning rate may hurt the shadow removal performance.

In addition, we visualise several samples of the learned lightness guidance and compared them with the restored images. We extract two kinds of feature maps from the first residual block R1 of LG-ShadowNet. One is taken from the input feature maps of R1, *i.e.*, the x_k^{Lab} in Eq. (9). The other is taken from the feature maps after multiplicative operation, *i.e.*, $x_k^{Lab} \odot x_m^L$ in Eq. (9) with $m = k$. The visualisation of these feature maps in terms of heat map are shown in Fig. 6. We observe that after the multiplicative operation, the activation values of non-shadow region decrease significantly, *e.g.*, the colour of each pixel presented in the heat map is close to blue, which means the model pays more attention on the shadow region. Especially from the second row of Fig. 6, we can see that the activation values of floor-tile joints in the non-shadow region are greatly suppressed.

Finally, while we propose to train Module I and Module II in LG-ShadowNet individually, we also try to train them jointly on the AISTD dataset. The jointly trained model achieves slightly worse results: it achieves 5.24, 4.18, 11.30 on the whole Lab image, non-shadow region, and shadow region in terms of RMSE, respectively. Note that in the training stage, the jointly trained model leads to a higher memory occupation and a longer training time.

TABLE VIII

QUANTITATIVE RESULTS ON ISTD AND AISTD IN TERMS OF RMSE, PSNR, AND SSIM. '-' DENOTES THE RESULT IS NOT PUBLICLY REPORTED. THE RESULTS OF THESE METHODS ARE EITHER OBTAINED FROM THE THEIR OFFICIAL RESULTS/PUBLICATIONS OR PRODUCED BY US USING THEIR OFFICIAL CODES (MARKED WITH '*')

Data Types	Methods	Training	ISTD					AISTD				
			Lab	N	S	PSNR	SSIM	*Lab*	N	S	PSNR	SSIM
Prior-based	Yang *et al.* [64]	N/A	15.63	14.83	19.82	20.11	0.655	16.74	15.09	24.36	19.77	0.662
	Guo *et al.* [8]	N/A	9.30	7.46	18.95	22.33	0.841	7.15	4.32	21.45	24.27	0.851
	Gong *et al.* [65]	N/A	8.53	7.29	14.98	24.07	0.908	5.10	3.39	14.42	27.78	0.916
Paired	ARGAN [11]	RGB	11.02	10.21	15.49	24.25	0.727	-	-	-	-	-
	ST-CGAN [14]	RGB	7.47	6.93	10.33	24.85	0.743	9.40	8.75	12.72	23.14	0.753
	DSC [12]	*Lab*	6.67	6.14	9.76	26.62	0.845	-	-	-	-	-
	SP+M-Net [18]	RGB	-	-	-	-	-	4.41	3.64	8.84	30.20	0.924
Unpaired	CycleGAN [16]	RGB	8.16	-	-	-	-	-	-	-	-	-
	Mask-ShadowGAN* [15]	RGB	7.41	6.68	12.67	24.63	0.821	5.48	4.52	11.53	27.65	0.918
	Mask-ShadowGAN* [15]	*Lab*	7.32	6.57	12.65	25.07	0.893	5.84	4.82	12.28	26.89	0.905
	LG-ShadowNet	*L*+*Lab*	**6.67**	**5.91**	**11.63**	**25.92**	**0.909**	**5.02**	**4.02**	**10.64**	**28.07**	**0.920**

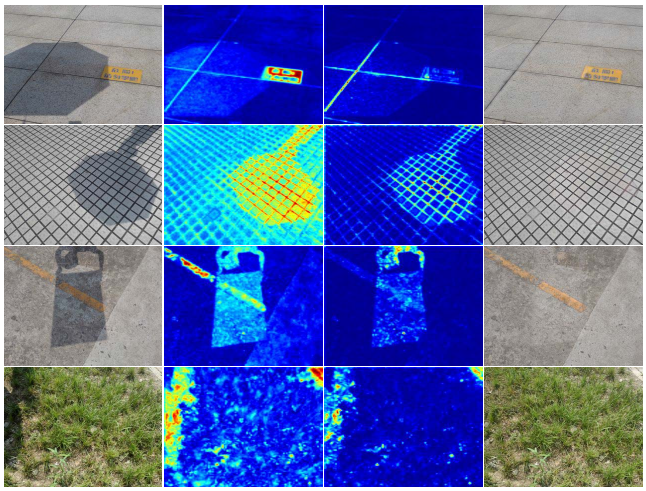


Fig. 6. Visualisation of the learned guidance on four samples from the AISTD testing set. x_k^{Lab} represents the input features of R1. $x_k^{Lab} \odot x_m^L$ represents the features after the multiplicative operation.

Fig. 7. Visual comparisons on USR. Each row shows results for one sample image.

E. Comparison With the State-of-the-Art

In this subsection, we compare our full model with several state-of-the-art methods on the ISTD, AISTD and USR datasets. Results of Mask-ShadowGAN are obtained by training and testing on each dataset using the code provided by its authors, while other results are provided by the authors of ST-CGAN [14], DSC [12] and SP+M-Net [18].

First of all, we compare our method with Mask-ShadowGAN on the USR dataset. The results of SSEQ, NIQE, and DBCNN are shown in Table VII. We observe that our method achieves the best performance in terms of NIQE, SSEQ, and DBCNN, which shows that the quality of shadow removal results can be improved by the proposed lightness-guided framework. For the user study, Mask-ShadowGAN trained on RGB data reports the most recent state-of-the-art performance. The proportions of votes received by LG-ShadowNet when compared with Mask-ShadowGAN trained on RGB and *Lab* data are 80.7% and 72.7%, respectively. These results show that, after converting the input data from RGB to *Lab*, Mask-ShadowGAN actually performs even better on USR. However, the proposed LG-ShadowNet still receives more votes than Mask-ShadowGAN trained on RGB or *Lab* data. Qualitative results are shown in Fig. 7.

Next, we compare the proposed method with the state-of-the-art methods on the ISTD and AISTD datasets. Among them, Guo *et al.* [8], Gong and Cosker [65], and Yang *et al.* [64] remove shadows based on image priors. ST-CGAN [14], DSC [12], and ARGAN [11] are trained using paired shadow and shadow-free images. SP+M-Net [18] is trained by using shadow and shadow-free image pairs, as well as shadow masks. CycleGAN [16] and Mask-ShadowGAN [15] are trained using unpaired images.

The quantitative results are shown in Table VIII. We can see that our method outperforms the methods based on image priors and those using unpaired data on both datasets. Compared with the methods using paired data, our method is also competitive, achieving comparable results to DSC [12] on the ISTD dataset. Note that Module II* trained on *Lab* data performs better than Mask-ShadowGAN on AISTD, and the former has fewer parameters. This proves the effectiveness of using the strategy of SqueezeNet [27] to reduce the model parameters. The comparison of the number of parameters, FLOPs, and

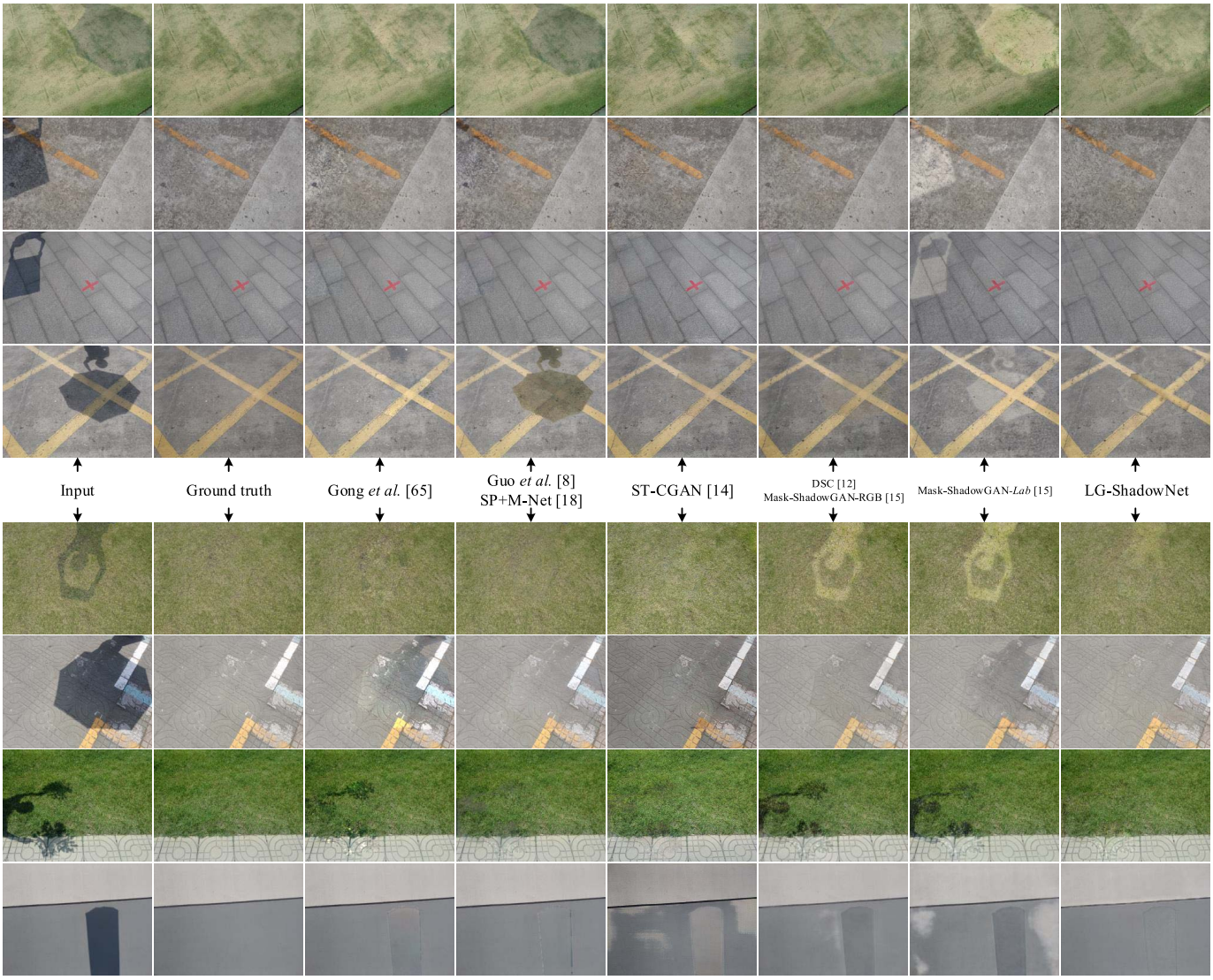


Fig. 8. Visual comparisons on ISTD and AISTD. Each row shows results for one sample image.

run time per image of the proposed LG-ShadowNet and the compared Mask-ShadowGAN [15] is shown in Table IX.

Figure 8 shows the qualitative results of LG-ShadowNet and several state-of-the-art methods on four challenging sample images in the ISTD (rows 1-4) and AISTD (rows 5-8) datasets. Compared with Mask-ShadowGAN, LG-ShadowNet restores the lightness of the shadow regions on all samples better and has less artefacts. However, the results of our method on some samples in the ISTD dataset (rows 1 and 3, the last column) look like bleaching artefacts and the colour is not well restored. Such bleaching artefacts are actually quite common in the results of other state-of-the-art methods [14], [15], [18]. One reason may be that our method pays more attention to the lightness and somehow ignores the restoration of other colours. A better balance between the use of the lightness information and the other colour information may alleviate this problem, which we will study in our future works.

Our method is also comparable to the methods using paired data, especially on the samples in ISTD. It is worth noting that our method can deal with the shadow edges better than SP+M-Net (rows 5-8 and column 4). The reason is that our method uses the continuous lightness information to guide the shadow removal while SP+M-Net uses binary shadow masks. These visual results verify the effectiveness of the proposed method for shadow removal.

TABLE IX

THE NUMBER OF PARAMETERS, FLOPS, AND RUN TIME PER IMAGE OF MASK-SHADOWGAN AND THE PROPOSED LG-SHADOWNET. THE RUN TIME PER IMAGE IS THE AVERAGE OF 100 RUNS. THE INPUT IMAGE SIZE IS 480 × 640

Methods	#Params	FLOPs	Run Time
Mask-ShadowGAN [15]	11.378M	266.4G	492ms
LG-ShadowNet (Ours)	3.535M	82.7G	242ms

Above all, our method exhibits the following advantages: 1) it is trained on unpaired data which are much easier to collect with a large variety in practice than paired data; 2) it has fewer parameters, lower complexity, and faster run time

per image than Mask-ShadowGAN; 3) it outperforms the state-of-the-art unpaired method Mask-ShadowGAN and traditional methods on public datasets.

V. Conclusion

In this paper, we proposed a new lightness-guided method for shadow removal using unpaired data. It fully explores the important lightness information by first training a CNN module only for lightness before considering other colour information. Another CNN module is then trained with the guidance of lightness information from the first CNN module to integrate the lightness and colour information for shadow removal. A colour loss is proposed to further utilise the colour prior of existing data. Experimental results verified the effectiveness of the proposed lightness-guided architecture and demonstrated that our LG-ShadowNet outperforms the state-of-the-art methods with training on unpaired data.

References

[1] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1337–1342, Oct. 2003.

[2] C. R. Jung, "Efficient background subtraction and shadow removal for monochromatic video sequences," *IEEE Trans. Multimedia*, vol. 11, no. 3, pp. 571–577, Apr. 2009.

[3] S. Nadimi and B. Bhanu, "Physical models for moving shadow and object detection in video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 1079–1087, Aug. 2004.

[4] A. Sanin, C. Sanderson, and B. C. Lovell, "Improved shadow removal for robust person tracking in surveillance scenarios," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 141–144.

[5] H. Liang, G. Liu, H. Zhang, and T. Huang, "Neural-network-based event-triggered adaptive control of nonaffine nonlinear multiagent systems with dynamic uncertainties," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 14, 2020, doi: 10.1109/TNNLS.2020.3003950.

[6] Y. Pan, P. Du, H. Xue, and H.-K. Lam, "Singularity-free fixed-time fuzzy control for robotic systems with user-defined performance," *IEEE Trans. Fuzzy Syst.*, early access, Jun. 3, 2020, doi: 10.1109/TFUZZ.2020.2999746.

[7] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew, "On the removal of shadows from images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 59–68, Jan. 2006.

[8] R. Guo, Q. Dai, and D. Hoiem, "Paired regions for shadow detection and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2956–2967, Dec. 2013.

[9] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri, "Automatic shadow detection and removal from a single image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 431–446, Mar. 2016.

[10] L. Zhang, Q. Zhang, and C. Xiao, "Shadow remover: Image shadow removal based on illumination recovering optimization," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4623–4636, Nov. 2015.

[11] B. Ding, C. Long, L. Zhang, and C. Xiao, "ARGAN: Attentive recurrent generative adversarial network for shadow detection and removal," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10213–10222.

[12] X. Hu, C.-W. Fu, L. Zhu, J. Qin, and P.-A. Heng, "Direction-aware spatial context features for shadow detection and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2795–2808, Nov. 2020.

[13] L. Qu, J. Tian, S. He, Y. Tang, and R. W. H. Lau, "DeshadowNet: A multi-context embedding deep network for shadow removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4067–4075.

[14] J. Wang, X. Li, and J. Yang, "Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1788–1797.

[15] X. Hu, Y. Jiang, C.-W. Fu, and P.-A. Heng, "Mask-ShadowGAN: Learning to remove shadows from unpaired data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2472–2481.

[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[17] G. D. Finlayson, S. D. Hordley, and M. S. Drew, "Removing shadows from images," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 823–836.

[18] H. Le and D. Samaras, "Shadow removal via shadow image decomposition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8578–8587.

[19] Y. Shor and D. Lischinski, "The shadow meets the mask: Pyramid-based shadow removal," *Comput. Graph. Forum*, vol. 27, no. 2, pp. 577–586, Apr. 2008.

[20] T. Chalidabhongse, D. Harwood, and L. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in *Proc. Int. Conf. Comput. Vis.*, 1999, pp. 1–19.

[21] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, and S. Sirotti, "Improving shadow suppression in moving object detection with HSV color information," in *Proc. IEEE Trans. Intell. Transp. Syst.*, Aug. 2001, pp. 334–339.

[22] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," 2017, *arXiv:1710.10196*. [Online]. Available: https://arxiv.org/abs/1710.10196

[23] H. Le, T. F. Y. Vicente, V. Nguyen, M. Hoai, and D. Samaras, "A+d net: Training a shadow detector with adversarial shadow attenuation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 662–678.

[24] M. Tkalcic and J. F. Tasic, "Colour spaces: Perceptual, historical and applicational background," in *Proc. IEEE Region 8 EUROCON Comput. Tool*, vol. 1, 2003, pp. 304–308.

[25] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4768–4777.

[26] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6849–6857.

[27] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, *arXiv:1602.07360*. [Online]. Available: http://arxiv.org/abs/1602.07360

[28] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inform. Process. Syst.*, 2014, pp. 2672–2680.

[29] M. Gryka, M. Terry, and G. J. Brostow, "Learning to remove soft shadows," *ACM Trans. Graph.*, vol. 34, no. 5, p. 153, 2015.

[30] C. Xiao, R. She, D. Xiao, and K.-L. Ma, "Fast shadow removal using adaptive multi-scale illumination transfer," *Comput. Graph. Forum*, vol. 32, no. 8, pp. 207–218, 2013.

[31] T. F. Y. Vicente, M. Hoai, and D. Samaras, "Leave-one-out kernel optimization for shadow detection and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 682–695, Mar. 2018.

[32] Y.-H. Lin, W.-C. Chen, and Y.-Y. Chuang, "BEDSR-net: A deep shadow removal network from a single document image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12905–12914.

[33] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.

[34] E. Gundogdu, V. Constantin, A. Seifoddini, M. Dang, M. Salzmann, and P. Fua, "GarNet: A two-stream network for fast and accurate 3D cloth draping," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8739–8748.

[35] P. Gupta and N. Rajput, "Two-stream emotion recognition for call center monitoring," in *Proc. 8th Annu. Conf. Int. Speech Commun. Assoc.*, 2007, pp. 1–4.

[36] X. Peng and C. Schmid, "Multi-region two-stream R-CNN for action detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 744–759.

[37] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12026–12035.

[38] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[39] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1831–1839.

[40] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1053–1061.

[41] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 2001, pp. 341–346.

[42] A. Levin, A. Zomet, and Y. Weiss, "Learning how to inpaint from global image statistics," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, p. 305.

[43] A. Telea, "An image inpainting technique based on the fast marching method," *J. Graph. Tools*, vol. 9, no. 1, pp. 23–34, Jan. 2004.

[44] M. M. Wilczkowiak, G. J. Brostow, B. Tordoff, and R. Cipolla, "Hole filling through photomontage," in *Proc. Brit. Mach. Vis. Conf.*, 2005, pp. 492–501.

[45] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[46] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Jul. 2017.

[47] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[48] W. Xiong *et al.*, "Foreground-aware image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5840–5848.

[49] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, "Contextual residual aggregation for ultra high-resolution image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7508–7517.

[50] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," 2016, *arXiv:1611.02200*. [Online]. Available: http://arxiv.org/abs/1611.02200

[51] S. Murali and V. Govindan, "Removal of shadows from a single image," in *Proc. Int. Conf. Futuristic Trends Comput. Sci. Eng.*, vol. 4, 2012, pp. 111–114.

[52] S. Murali and V. K. Govindan, "Shadow detection and removal from a single image using LAB color space," *Cybern. Inf. Technol.*, vol. 13, no. 1, pp. 95–103, Mar. 2013.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[54] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[55] S. Gross and M. Wilber, "Training and investigating residual nets," Facebook AI Res., Menlo Park, CA, USA, Tech. Rep., 2016. [Online]. Available: http://torch.ch/blog/2016/02/04/resnets.html

[56] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[57] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022

[58] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Process., Image Commun.*, vol. 29, no. 8, pp. 856–863, Sep. 2014.

[59] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Nov. 2012.

[60] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[62] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 261–270.

[63] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[64] Q. Yang, K.-H. Tan, and N. Ahuja, "Shadow removal using bilateral filtering," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4361–4368, Oct. 2012.

[65] H. Gong and D. Cosker, "Interactive shadow removal and ground truth for variable scene categories," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–11.

Zhihao Liu received the B.E. degree from the Beijing Institute of Graphic Communication, Beijing, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing. His current research interests include computer vision and pattern recognition.

Hui Yin received the Ph.D. degree in computer application technology from Beijing Jiaotong University, Beijing, China. She is currently a Full Professor with the School of Computer and Information Technology, Beijing Jiaotong University. Her current research interests include the machine vision, intelligent information processing, and their application in the railway industry.

Yang Mi received the B.S. degree in information engineering from the Beijing Institute of Technology, China, in 2009, and the M.S. degree and the Ph.D. degree in computer science and engineering from the University of South Carolina in 2018 and 2020, respectively. He is currently an Assistant Professor with the Department of Data Science and Engineering, China Agriculture University. His research interests include computer vision and machine learning.

Mengyang Pu is currently pursuing the Ph.D. degree with the Department of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. Her research interests include computer vision, image segmentation, and edge detection.

Song Wang (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign (UIUC) in 2002. From 1998 to 2002, he worked as a Research Assistant with the Image Formation and Processing Group, Beckman Institute, UIUC. In 2002, he joined the Department of Computer Science and Engineering, University of South Carolina, where he is currently a Professor. His research interests include computer vision, medical image processing, and machine learning. He is a Senior Member of the IEEE Computer Society. He is also serving as an Associate Editor for the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Pattern Recognition Letters*, and *Electronics Letters*.