Structure-Informed Shadow Removal Networks

Yuhao Liu, Qing Guo, *Member, IEEE*, Lan Fu, Zhanghan Ke, *Graduate Student Member, IEEE*, Ke Xu, Wei Feng, *Member, IEEE*, Ivor W. Tsang, *Fellow, IEEE*, and Rynson W. H. Lau, *Senior Member, IEEE*

***Abstract*— Existing deep learning-based shadow removal methods still produce images with shadow remnants. These shadow remnants typically exist in homogeneous regions with low-intensity values, making them untraceable in the existing image-to-image mapping paradigm. We observe that shadows mainly degrade images at the image-structure level (in which humans perceive object shapes and continuous colors). Hence, in this paper, we propose to remove shadows at the image structure level. Based on this idea, we propose a novel structure-informed shadow removal network (*StructNet*) to leverage the image-structure information to address the shadow remnant problem. Specifically, StructNet first reconstructs the structure information of the input image without shadows and then uses the restored shadow-free structure prior to guiding the image-level shadow removal. StructNet contains two main novel modules: 1) a *mask-guided shadow-free extraction (MSFE) module* to extract image structural features in a non-shadow-to-shadow directional manner; and 2) a *multi-scale feature & residual aggregation (MFRA) module* to leverage the shadow-free structure information to regularize feature consistency. In addition, we also propose to extend *StructNet* to exploit multi-level structure information (*MStructNet*), to further boost the shadow removal performance with minimum computational overheads. Extensive experiments on three shadow removal benchmarks demonstrate that our method outperforms existing shadow removal methods, and our StructNet can be integrated with existing methods to improve them further.**

***Index Terms*— Single-image shadow removal, image structure, structure-level shadow removal.**

I. Introduction

SHADOWS exist everywhere. They appear on surfaces where light cannot reach due to occlusions. Faithfully recovering the original color and textures of shadow regions helps facilitate many other tasks, *e.g.*, light source analysis [1], face recognition [2], object detection [3], and novel image creation [4]. Hence, shadow removal is a long-standing problem in computer vision and graphics, with many methods proposed.

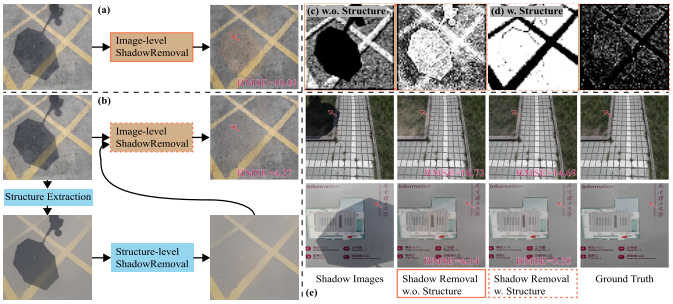


Fig. 1. (a) State-of-the-art shadow removal methods (*e.g.*, AEF [8]) typically learn a direct shadow-to-shadow-free mapping and may often produce shadow remnants with color artifacts. (b) We propose to incorporate image-structure information into the shadow removal process. We visualize the features of approaches (a) and (b) in (c) and (d), respectively, which show that features of (d) are structured according to region homogeneity. (e) Results of original AEF and its structure-enhanced counterpart, where red arrows indicate the region with shadow remnants exist, and RMSE metric are shown for reference.

Conventional shadow removal methods are typically based on modeling varied intensity [5] and illumination [6], or involving user interaction [7]. They usually fail when the prior assumptions are not satisfied or the scenes are intricate.

Deep learning-based shadow removal methods [8], [9], [10], [11], [12] achieve impressive performances in recent years due to the high generalization capability of advanced neural networks. These methods typically formulate the shadow removal problem as a shadow-to-shadow-free images mapping. Qu et al. [9] first use CNNs to extract shadow-related information (*i.e.*, location, appearance, and semantic information) and then predict the shadow matte for shadow removal. Fu et al. [8] use CNNs to predict exposure parameters and then remove shadows by fusing multiple shadow exposures. However, these state-of-the-art methods may still produce unsatisfactory results with shadow remnants and color artifacts. In Fig. 1(a), we can see yellowish shadow remnants in the result from AEF [8]. These remnants are usually internally homogeneous and of low intensity values, making them hard to detect by the existing image-level shadow removal paradigm represented by [8].

In this work, we propose to address the shadow remnant problem by incorporating the image-structure information (which consists of low-frequency image components that represent the object colors and shapes), as shown in Fig. 1(b)). While the structure layer of an image is the primary information perceived by the human vision system [13], [14], it separates the observed objects into multiple homogeneous regions with similar colors and intensities [15]. Hence, it

Manuscript received 25 March 2023; revised 24 August 2023 and 30 September 2023; accepted 3 October 2023. Date of publication 17 October 2023; date of current version 1 November 2023. This work was supported in part by the Strategic Research Grant (SRG) Grants from the City University of Hong Kong under Grant 7005674 and Grant 7005843; in part by the National Research Foundation, Singapore, and Defence Science Organisation (DSO) National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-GC-2023-008). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nikos Deligiannis. *(Yuhao Liu and Qing Guo are co-first authors.) (Corresponding authors: Qing Guo; Rynson W. H. Lau.)*

Yuhao Liu, Zhanghan Ke, Ke Xu, and Rynson W. H. Lau are with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: yuhaoliu7456@outlook.com; zhanghake2-c@my.cityu.edu.hk; kkangwing@gmail.com; Rynson.Lau@cityu.edu.hk).

Qing Guo and Ivor W. Tsang are with the Institute of High Performance Computing (IHPC) and the Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: tsingqguo@ieee.org; ivor_tsang@ihpc.a-star.edu.sg).

Lan Fu is with InnoPeak Technology Inc., Palo Alto, CA 94303 USA (e-mail: lan.fu@innopeaktech.com).

Wei Feng is with the College of Intelligence and Computing, Tianjin University, Tianjin 300072, China (e-mail: wfeng@ieee.org).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2023.3323814, provided by the authors.

Digital Object Identifier 10.1109/TIP.2023.3323814

should be much easier to locate and much cleaner to remove shadows in the image-structure layer, due to the absence of high-frequency texture details. With the recovered shadow-free image-structure layer as guidance, it may then be possible to restore object details in shadow regions.

To verify our idea, we use the naive UNet [16] to first perform image-structure shadow removal, the output of which is then used to guide the image shadow removal process (Fig. 1(b)). With this model, we show that structure-level shadow removal can help boost the performances of a state-of-the-art shadow removal method [8] (Fig. 1(a) *vs* (b) and Fig. 1(e) column 2 *vs* 3). We visualize the feature maps of original AEF in Fig. 1(b) and the structure-enhanced counterpart in Fig. 1(c) and Fig. 1(d), respectively. We can see that features in (d) are structured based on region homogeneity, which helps alleviate the color artifacts of Fig. 1(a). However, we also note that the standard convolution used in the naive model (as well as in almost all existing methods) adopts spatially-shared weights to process both shadow and non-shadow regions, and neglects their distinct patterns, resulting in color shifts.

Based on the above analysis, we propose the *structure-informed shadow removal network (StructNet)*, which consists of the structure-level shadow removal step in stage-1 and the image-level shadow removal step in stage-2. We propose two novel modules to facilitate the shadow removal in the structure-level: *mask-guided shadow-free extraction (MSFE)* and *multi-scale feature & residual aggregation (MFRA) modules*. The MSFE module aims to model non-shadow-to-shadow structure information conditioned on the non-shadow regions, while the MFRA module focuses on incorporating the extracted shadow-free structure information into the shadow removal process with feature consistency regularization. They can dynamically extract shadow-free structure information and propagates them into shadow regions for shadow removal. We conduct extensive experiments on three benchmarks to evaluate our method and show that StructNet outperforms state-of-the-art shadow removal methods. StructNet can also be incorporated into existing fully-supervised shadow removal methods to help enhance their performances. Finally, we propose to conduct the shadow removal task at multiple structure levels with a single architecture (named MStructNet), which is not only efficient but also outperforms state-of-the-art methods. In summary, we make the following efforts:

- We construct a naive model (*i.e.*, the vanilla UNet) for structure-level shadow removal and conduct extensive empirical studies on it. We show that removing shadows at the structure level is more effective than that at the image level, and the restored shadow-free structures can improve the quality of the output images.
- We propose the *structure-informed shadow removal network (StructNet)*, which contains two novel modules for structure-level shadow removal: mask-guided shadow-free extraction (MSFE) module and multi-scale feature & residual aggregation (MFRA) module. MSFE learns directional shadow-free structure information from non-shadow to shadow regions, while MFRA regularizes feature consistency by dynamically fusing the output from MSFE with whole image features.
- We further propose a self-contained shadow removal method, multi-level StructNet (MStructNet), which utilizes multi-level shadow structures at the feature level with low parameters for high-quality shadow removal.
- Extensive evaluations and ablation studies on three public datasets show that the proposed StructNet can help enhance the performances of existing SOTA methods, and MStructNet achieves high-quality image restoration, outperforming SOTA shadow removal methods.

II. Related Work

A. Shadow Removal

Traditional-based shadow removal methods [17], [18], [19], [20], [21], [22], [23] mainly rely on image statistical priors (*e.g.*, gradients and colors). Finlayson et al. [19], [24] solve shadow detection and removal via gradient consistency of illumination invariation. Shor and Lischinki [25] propose an illumination-based model in which a pixel-wise relationship between shadow and shadow-free pixel intensities is modeled. Guo et al. [6] propose a relative illumination model based on paired data modeling. However, conventional methods often fail when their hand-crafted features do not represent real-world scenes.

Deep learning-based techniques, renowned for their advanced modeling capabilities, have found extensive applications in various vision tasks such as detection [26], segmentation [27], [28], and generation [29], [30]. With the availability of large-scale datasets in shadow removal [9], [11], numerous approaches [8], [10], [11], [31], [32], [33], [34], [35], [36] have been proposed. Typically, these methods model shadow removal as an image-to-image mapping process from shadow image to shadow-free image. DeShadowNet [9] first proposes to use multi-branch CNNs to extract multi-level contexts for shadow removal. The follow-ups focused on modeling the shadow formation model [37], [38], and designing different network architectures and exploiting distinctive properties (*e.g.*, contexts [10], exposures [8], residuals [32], and illuminations [39]). Unpaired/unsupervised methods [12], [40], [41], [42], [43], [44] have also been proposed to alleviate the labeling cost of paired data through generative adversarial training and pseudo labels generation. Nonetheless, these methods may still produce shadow remnants and color artifacts. In this paper, we propose to model the image structure layer to handle the shadow remnant problem. Naoto et al. [45] further propose to generate a synthetic shadow dataset for shadow removal.

B. Image-Structure in Vision Tasks

The image-structure information [13], [14] has been studied in several vision tasks. Ren et al. [46] propose to leverage the image-structure information to guide their inpainting method to generate image content in a low-to-high frequency manner. Gui et al. [47] propose to leverage the intermediate image-structure layers to constrain the smoothness of consecutive frames for video interpolation. For cartoonization, Wang et al. [48] propose to process the image-structure layer separately from the texture layer to maintain harmonious colors.

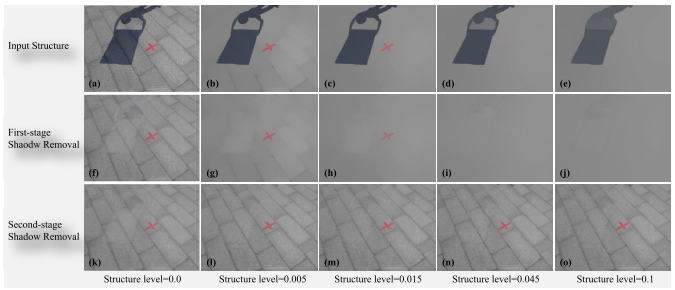


Fig. 2. Shadow removal results at different structure levels. The 1st row shows the original shadow image (a) and its structures (b)-(e) extracted by [49] at four different structure levels (*i.e.*, $l \in \{0.005, 0.015, 0.045, 0.1\}$). The 2nd row shows the shadow removal results by feeding the shadow structures in the 1st row to respective vanilla UNets. Image (f) represents the result of the image-level shadow removal, while images (g)-(j) are the results of structure-level shadow removal with $l > 0.0$. The 3rd row shows restoration results of our naive two-stage shadow removal network by feeding the restored shadow-free structures (*i.e.*, the images at 2nd row) into the second vanilla UNets.

In this paper, we leverage the image-structure information to locate and track the shadow remnants in homogeneous regions to remove shadows and preserve color consistency.

III. Structure-Level Shadow Removal

In this section, we introduce our structure-level shadow removal approach. We first formulate the structure-level shadow removal problem in Section III-A and then investigate the application of structure information in shadow removal in Section III-B.

A. Formulation of Structure-Level Shadow Removal

In structure-level shadow removal, we first use a structure extraction method $\varphi(\cdot)$ to map the shadow image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ to its structure image/layer, in which image inherent colors and main outlines are preserved while detailed textures are removed (see Fig. 1(b) and Fig. 2), as

$$\mathbf{S}_l = \varphi(\mathbf{I}, l), \tag{1}$$

where $l > 0.0$ is a hyper-parameter determining the structure level, and $\mathbf{S}_l \in \mathbb{R}^{H\times W\times 3}$ is the structure image at the lth structure level. A higher l will remove more detailed textures (see the first row of Fig. 2). We follow the setups in [8] and [37] to formulate the lth structure-level shadow removal:

$$\hat{\mathbf{S}}_l = \phi_l(\mathbf{S}_l, \mathbf{M}), \tag{2}$$

where $\phi_l(\cdot)$ is the shadow removal model corresponding to $\mathbf{S}_l$, and $\mathbf{M} \in \mathbb{R}^{H\times W}$ is a binary mask that indicates shadow and non-shadow pixels with 1 and 0, respectively. Note that the shadow mask is an input to the shadow removal task. The output $\hat{\mathbf{S}}_l$ is the restored structure layer at the lth structure level, *i.e.*, the result of structure-level shadow removal.

B. Empirical Studies

To study how the structure information affects shadow removal results, we employ the structure extraction model proposed by Xu et al. [49] as $\varphi(\cdot)$. We design a variant of vanilla UNet [16], which consists of an encoder with 5 convolution layers and a decoder with 5 de-convolution layers, as $\phi_l(\cdot)$.

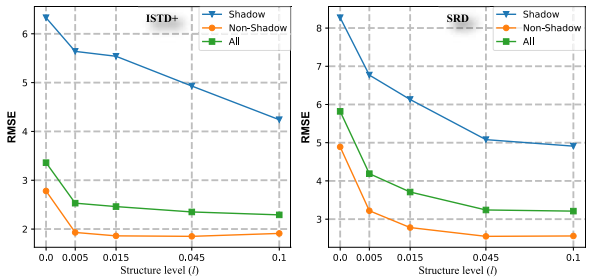


Fig. 3. Comparison of the image-level (*i.e.*, $l = 0.0$) and four structure-level shadow removal process with $l \in \{0.005, 0.015, 0.045, 0.1\}$ on two public datasets (*i.e.*, ISTD+ [38] and SRD [9]). We employ the root mean square error (RMSE) in the LAB color space as the evaluation metric to assess the shadow-removal performances in the shadow regions, non-shadow regions, and the whole (*i.e.*, All) image, respectively.

Each layer in our $\phi_l(\cdot)$ is followed by an Instance Norm [50] function and a Leaky-ReLU [51] (for the encoder) or ReLU (for the decoder) function. We set the kernel size, padding, and stride of each layer to 4, 2, and 1, respectively. Based on the above network configurations, we aim to answer the following three questions: ❶ how does the capability of shadow removal vary at different structure levels? ❷ whether the structure-level shadow removal results (*i.e.*, corrected structure) could guide the image-level shadow removal? ❸ whether existing model architectures are suitable for structure-level shadow removal?

1) Shadow Removal at Different Structure Levels: Since the shadow removal results may vary at different structure levels (lth), we train and test $\phi_l(\cdot)$ at five structure levels $l \in \{0.0, 0.005, 0.015, 0.045, 0.1\}$ [1] Note that $l = 0.0$ is equivalent to image-level shadow removal, *i.e.*, Eq. 1 with $l = 0.0$ as an identity function. To avoid the possible influence of elaborately designed loss functions, we only optimize the prediction $\hat{\mathbf{S}}_l = \phi_l(\varphi(\mathbf{I}, l), \mathbf{M})$ via the mean absolute error $L_1(\hat{\mathbf{S}}_l, \mathbf{S}_l^*) = \|\hat{\mathbf{S}}_l - \mathbf{S}_l^*\|_1$, where $\mathbf{S}_l^* = \varphi(\mathbf{I}^*, l)$ is the ground truth structure generated from the shadow-free image $\mathbf{I}^*$. On the validation set, we calculate the root mean square error (RMSE) between $\hat{\mathbf{S}}_l$ and $\mathbf{S}_l^*$ after converting them into the LAB color space. The smaller, the better.

We conduct evaluations on two widely used datasets, ISTD+ [38] and SRD [9]. Based on the results shown in Fig. 3, we observe that ❶ the RMSE on shadow regions decreases continuously as l increases. This suggests that it is easier to obtain high quality shadow removal results at the structure level (*i.e.*, $l > 0$) than at the image level (*i.e.*, $l = 0$). Such a phenomenon is also reflected in visual results shown in Fig. 2, in which there are obvious artifacts in the image-level shadow removal result (Fig. 2(f)), but such artifacts are greatly reduced at the structure-level shadow removal results (Fig. 2(g)-(j)). ❷ The RMSE curves in the non-shadow regions descend at the beginning then become flat when l increases to reach a certain level. The RMSE curves of the whole images have similar shapes to those of non-shadow regions. For non-shadow regions of the ISTD+ dataset, $l = 0.1$ has even worse RMSE than that of $l = 0.045$ (Fig. 3 left).

These experiments show that a higher structure level l generally facilitates shadow removal by making the shadow removal network focus more on color and structure information instead

[1]The range of structure levels are determined by the structure extraction method. When the level is larger than 0.1, the image will degrade to a pure color map.

TABLE I

COMPARISON BETWEEN DIRECT SINGLE-STAGE IMAGE-LEVEL SHADOW REMOVAL AND FIVE VARIANTS ($l = 0.0$ IS REGARDED AS A SPECIAL VARIANT.) OF TWO-STAGE STRUCTURE-LEVEL SHADOW REMOVAL. ALL EXPERIMENTS ARE CONDUCTED ON THE ISTD+ DATASET WITH THE VANILLA UNET AND L_1 LOSS FUNCTION

	Structure level l for the first stage	Shadow ↓	Non-shadow ↓	All ↓
Two-stage shadow removal	0.0	6.28	2.99	3.53
	0.005	5.98	2.55	3.11
	0.015	5.89	2.49	3.05
	0.045	6.17	2.57	3.16
	0.1	6.15	2.56	3.15
Direct Single-stage Image-level shadow removal		6.33	2.78	3.36

of texture information. However, if l is too large, it may lead to shadow spreading, *i.e.*, similar shadow visual patterns may appear in neighboring non-shadow regions (see Fig. 2(e)), which in turn causes a higher error in the non-shadow regions.

2) Shadow Removal With Structure-Level Guidance: We investigate if structure-level shadow removal is beneficial to image-level shadow removal and formulate a two-stage pipeline of which the first stage combines Eq. 1 and 2 for structure-level shadow removal. The second stage uses a new model $\psi_l(\cdot)$, which takes the corrected structure/layer $\hat{\mathbf{S}}_l$ as ancillary input for image-level shadow removal, as:

$$\hat{\mathbf{I}}_l = \psi_l(\mathbf{I}, \hat{\mathbf{S}}_l, \mathbf{M}), \tag{3}$$

where $\hat{\mathbf{I}}_l$ denotes the image-level shadow removal results guided by $\hat{\mathbf{S}}_l$. Theoretically, ψ_l can be an arbitrary image-level shadow removal method (*e.g.*, ST-CGAN [11] or AEF [8]). For simplicity, we simply assume ψ_l to have the same architecture as ϕ_l. When training the pipeline corresponding to $l \in \{0.0, 0.005, 0.015, 0.045, 0.1\}$, we fix ϕ_l optimized in Sec. III-B1 and learn ψ_l. We apply the same L_1 loss function and RMSE metric as in Sec. III-B1. Table I shows the results (on the ISTD+ dataset) of the two-stage shadow removal pipeline with different l and a single-stage image-level shadow removal model. Note that the pipeline with $l = 0.0$ can be regarded as a stack of two vanilla UNet models for image-level shadow removal. We observe that ❶ two-stage image-level (*i.e.*, $l = 0.0$) shadow removal does not yield better performance, compared to single-stage image-level shadow removal, and shadow remnants cannot be eliminated by simply adding more CNN parameters as shown in Fig. 2(a,f,k); ❷ two-stage shadow removal with $l > 0.0$ achieves lower RMSE than that of image-level shadow removal (either two-stage shadow removal with $l = 0.0$ or direct single-stage shadow removal), which shows that the restored shadow-free structures can help image-level shadow removal. We also observe from the results in Fig. 2 that the artifacts in Fig. 2(f) (*i.e.*, the result of single-stage image-level shadow removal) are alleviated by the two-stage shadow removal with $l > 0.0$, as shown in Fig. 2(l-o)).

3) Limitations of Using the Vanilla UNet: In Section III-B2, we have demonstrated that the structure-level shadow removal results can benefit image-level shadow removal to some degree. Here, we would like to know if the vanilla UNet is good enough for this two-stage shadow removal (*i.e.*, first

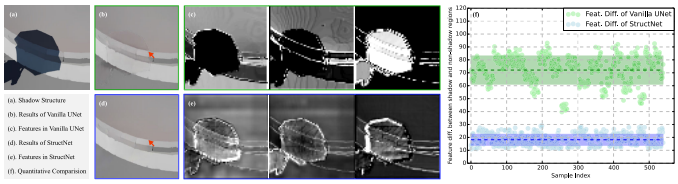


Fig. 4. Visualization and quantitative comparison of vanilla UNet and StructNet for structure-level shadow removal. (a) is the input shadow structure image, which is fed to the vanilla UNet and StructNet to obtain (b) and (d), respectively. Images (c) and (e) show the randomly sampled three feature channels produced by the 2nd convolutional layer of the two networks. In addition, we also extract the features from the 2nd convolution layer of the vanilla UNet and StructNet of all images in the ISTD+ test set. For each image, we calculate the absolute difference between the shadow and non-shadow regions in each feature channel and obtain the average difference across all channels. Image (f) shows the average feature differences of all images using the vanilla UNet (green points) and StructNet (blue points).

at structure-level and then at image-level). We observe that the standard convolution operations used in the vanilla UNet process shadow and non-shadow regions uniformly, and ignore the distinctions between them (*e.g.*, color-bias). In other words, the standard convolution used in the vanilla UNet attempts to map shadow and non-shadow regions that have very different appearances to the same pattern, which makes the learning of the convolution weights challenging. As a result, the vanilla UNet may produce obvious color shifts between shadow and non-shadow regions in the output image, as shown in Fig. 4(b).

To support the above analysis, we visualize three randomly selected feature channels of the 2nd convolution layer [2] in the vanilla UNet in Fig. 4(c). We can see that the features of the shadow and non-shadow regions show obvious divergences, although we expect them to be consistent in order to recover the colors of the shadow regions. We further conduct a quantitative analysis on the test set of ISTD+. For each sample, we first extract the features output by the 2nd convolution layer of the vanilla UNet. We then compute the means of it feature maps in the shadow and non-shadow regions separately and show the absolute difference between the two with a single point in Fig. 4(f). We can see that there are huge differences between shadow and non-shadow regions in the feature space. Such feature differences are caused by the uniform processing of standard convolutions used in the vanilla UNet. As a result, the vanilla UNet produces results with color shift. This motivates us to design a novel solution to overcome the problems of applying the vanilla UNet to structure-level shadow removal.

IV. STRUCTNET

In this section, we propose a novel two-stage model, named structure-informed shadow removal network (StructNet), to better utilize the structure-level shadow removal results (*i.e.*, the corrected structure image/layer) to guide the image-level shadow removal step. StructNet contains two novel designs: a mask-guided shadow-free extraction (MSFE) module in Section IV-A, and a multi-scale feature & residual aggregation (MFRA) module in Section IV-B. The configuration details of StructNet are then described in Section IV-C.

As outlined in Section III-B3, the standard convolution operations in vanilla UNet treat shadow and non-shadow regions

[2]The difference between shadow and non-shadow regions in deeper layers is minimal and indistinguishable. Thus, we choose the 2nd conv.

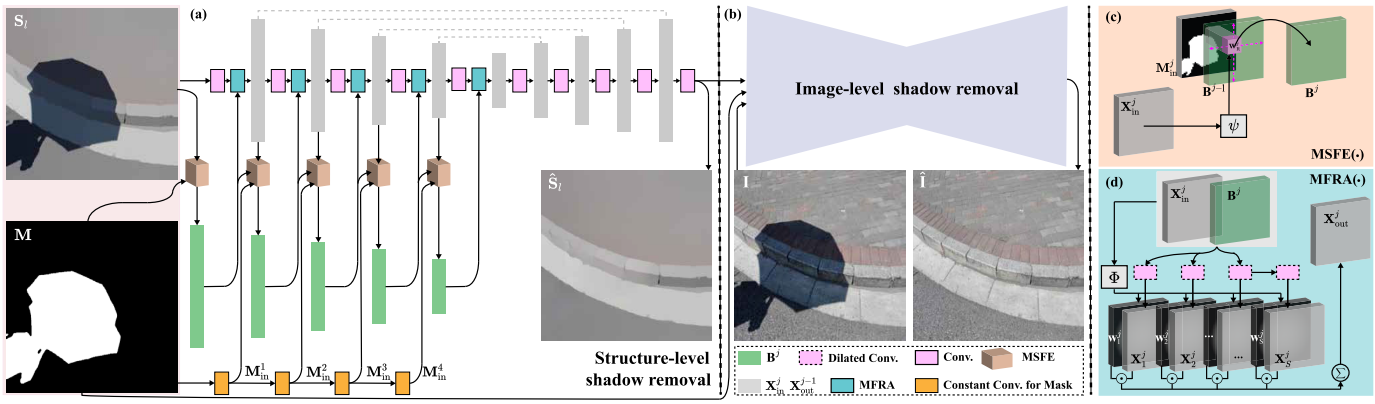


Fig. 5. Pipeline of the proposed StructNet. (a) shows the structure-level shadow removal. (b) shows the image-level shadow removal with the assistance of predicted shadow-free structure from (a). (c) and (d) represent the mask-guided shadow-free extraction (MSFE) and the multi-scale feature & residual aggregation (MFRA) modules, respectively, in the architecture.

uniformly, ignoring the differences between the two regions. Specifically, there is a noticeable feature shifting within the shadow region relative to the non-shadow region. The standard convolution in vanilla UNet seeks to map shadow and non-shadow regions, which exhibit highly dissimilar appearances, to an identical pattern, which is challenging. For instance, the features from the trained vanilla UNet manifest distinct appearances, as demonstrated in Fig. 4 (c). This discrepancy leads the vanilla UNet to generate evident color shifts between shadow and non-shadow regions in the output image, as depicted in Fig. 4 (b). To rectify the shortcomings of the standard convolution, we propose to make it shadow-aware. We introduce the addition of a directional bridge to the conventional convolution operations, which is guided by the non-shadow regions. This innovative approach promotes homogeneity between the shadow and non-shadow features, thus addressing the issues inherent in the previous method. Specifically, given the input features $\mathbf{X}_{\text{in}}^{j} \in \mathbb{R}^{H_{\text{in}}^{j} \times W_{\text{in}}^{j} \times C_{\text{in}}^{j}}$ at the jth layer, we propose to process the features as:

$$\mathbf{X}_{\text{out}}^{j} = \text{Fusion}(\mathbf{X}_{\text{in}}^{j} * \mathbf{W}^{j}, \mathbf{B}^{j}), \tag{4}$$

where $\mathbf{X}_{\text{out}}^{j} \in \mathbb{R}^{H_{\text{out}}^{j} \times W_{\text{out}}^{j} \times C_{\text{out}}^{j}}$ are the output features, $\mathbf{W}^{j}$ are the learnable weights, and $\mathbf{B}^{j} \in \mathbb{R}^{H_{\text{out}}^{j} \times W_{\text{out}}^{j} \times C_{\text{out}}^{j}}$ is a learned feature shifting tensor aiming to reduce the feature difference between the non-shadow and shadow regions. Fusion$(\cdot)$ is a function to fuse the shifting information in $\mathbf{B}^{j}$ and the features $\mathbf{X}_{\text{in}}^{j} * \mathbf{W}^{j}$ effectively, thus regularizing the output features to be consistent between the shadow and non-shadow regions. $\mathbf{B}^{j}$ is computed by Bridge$(\cdot)$, as:

$$\mathbf{B}^{j} = \text{Bridge}(\mathbf{X}_{\text{in}}^{j}, \mathbf{B}^{j-1}, \mathbf{M}_{\text{in}}^{j}), \tag{5}$$

where $\mathbf{M}_{\text{in}}^{j} \in \mathbb{R}^{H_{\text{in}}^{j} \times W_{\text{in}}^{j}}$ is a binary map that indicates the shadow regions with 1's and non-shadow regions with 0's. $\mathbf{B}^{j-1}$ is the shifting tensor of the previous layer, and Bridge$(\cdot)$ is trained to extract feature shifting of the non-shadow region shadow regions at the jth layer. Note that such a solution has two benefits: ❶ The advantages of the standard convolution are preserved via Eq. (4), which can extract perception across the whole image; ❷ The potential shifting between shadow and non-shadow regions is supplemented via Eq. (5).

With the above formulation, we propose the structure-informed shadow removal network (StructNet), as shown in Fig. 5. StructNet consists of two stages. The first stage performs structure-level shadow removal, while the second stage conducts image-level shadow removal guided by the results from the first stage. In the first stage, we propose the two novel modules, *i.e.*, MSFE and MFRA, to extensively exploit the structure information. The second stage can be any existing supervised shadow removal method.

A. The MSFE Module

Inspired by the segmentation-aware convolution [52], we propose to embed the shadow mask in the convolution operation explicitly and formulate Bridge$(\cdot)$ as:

$$\mathbf{B}^{j}[\mathbf{p}] = \alpha_{\mathbf{p}} \sum_{\mathbf{q} \in \mathcal{N}_{\mathbf{p}}} \mathbf{B}^{j-1}[\mathbf{q}](1 - \mathbf{M}_{\text{in}}^{j}[\mathbf{q}])\mathbf{W}_{\text{B}}^{j}[\mathbf{q} - \mathbf{p}], \tag{6}$$

where $\mathbf{W}_{\text{B}}^{j} \in \mathbb{R}^{K^j \times K^j \times C_{\text{in}}^{j} \times C_{\text{out}}^{j}}$ are the weights of a convolution layer, $\mathbf{p}$ and $\mathbf{q}$ are the coordinates of elements in $\mathbf{X}_{\text{in}}^{j}$, $\mathbf{M}^{j}$, $\mathbf{B}^{j}$, or $\mathbf{W}_{\text{B}}^{j}$. The set $\mathcal{N}_{\mathbf{p}}$ contains neighboring elements of $\mathbf{p}$, and its size is equal to the kernel size of $\mathbf{W}_{\text{B}}^{j}$ (*i.e.*, K^2). The normalization term $\alpha_{\mathbf{p}}$ is defined as $\frac{1}{\sum_{\mathbf{q} \in \mathcal{N}_{\mathbf{p}}} \mathbf{M}_{\text{in}}^{j}[\mathbf{q}]}$. The mask (*i.e.*, $\mathbf{M}_{\text{in}}^{j}$) is obtained by convoluting the mask from the previous layer (*i.e.*, $\mathbf{M}_{\text{in}}^{j-1}$) with a constant weight (*i.e.*, $\mathbf{W}_{\mathbb{1}}$ whose elements are one) through $\mathbf{M}_{\text{in}}^{j} = \mathbf{M}_{\text{in}}^{j-1} * \mathbf{W}_{\mathbb{1}}^{j}$. Intuitively, with Eq. 6, the output $\mathbf{B}^{j}$ only rely on the non-shadow regions due to the restriction of the mask $\mathbf{M}^{j}$ and can fill the gap across shadow and non-shadow regions.

As shown in Fig. 6, due to the low color intensity of the shadow regions, $\mathbf{X}_{\text{in}}^{j} * \mathbf{W}^{j}$ presents clearer shadow regions and focuses on brighter colors in the non-shadow regions. Instead, $\mathbf{B}^{j}$ can favorably attend to the shadow regions as the shadow mask introduces the positional information. In addition, we also show in column 4 the visualization of the features after adding the standard convolution and shift information (*i.e.*, $\mathbf{X}_{\text{in}}^{j} * \mathbf{W}^{j} + \mathbf{B}^{j}$). The results show that the direct addition of the two fails to achieve feature consistency. This is caused by the fact that each pixel strictly depends on the same position but ignores the convolutional perception and the non-shadow context of the shifting bridge. Finally, integrating the global perception $\mathbf{X}_{\text{in}}^{j} * \mathbf{W}^{j}$ and the offset $\mathbf{B}^{j}$, the result $\mathbf{X}_{out}^{j}$ exhibits consistent homogeneity across the shadow and non-shadow

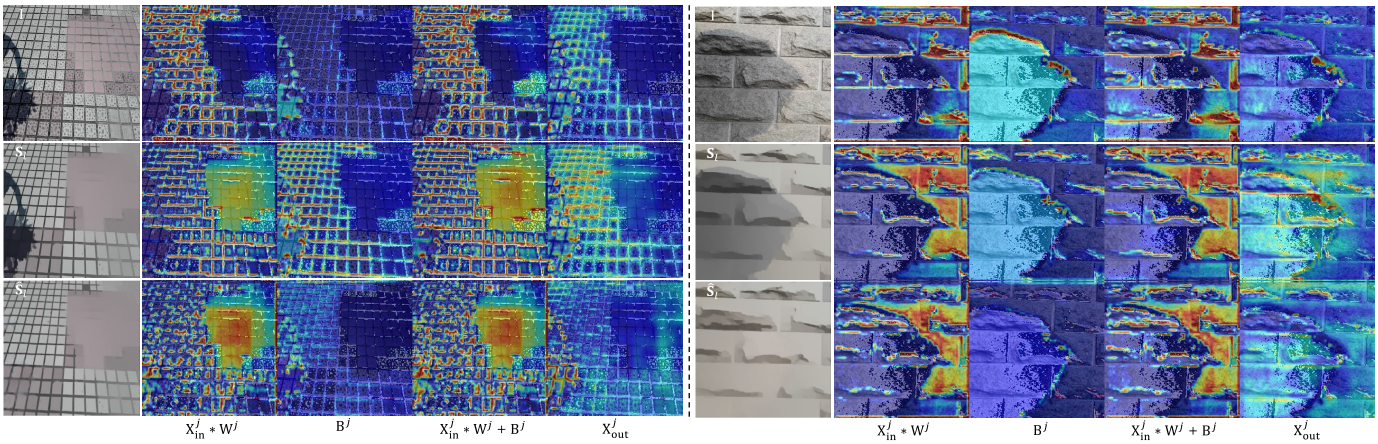


Fig. 6. Feature visualization of the global perception ($\mathbf{X}^j_{\text{in}} * \mathbf{W}^j$), the offset ($\mathbf{B}^j$), direct addition of $\mathbf{X}^j_{\text{in}} * \mathbf{W}^j$ and $\mathbf{B}^j$, and output features ($\mathbf{X}^j_{\text{out}}$) by fusing the former two with MFRA.

regions. In addition, instead of training the weight $\mathbf{W}^j_{\text{B}}$ for all examples, we propose to make it dynamically modulated according to different input features, *i.e.*, $\mathbf{W}^j_{\text{B}} = \eta(\mathbf{X}^j_{\text{in}})$, where $\eta(\cdot)$ is a sub-network having two convolution layers.

Our MSFE is different from the partial convolution [53] in two aspects: ❶ The convolution weights of the proposed bridge function are conditional on the input whole scene features, while those of the partial convolution is fixed after training; ❷ The operations with Eq. 4 and Eq. 5 are a combination of standard and dynamic partial convolution. The former aims to extract the perception of the whole image, while the latter is to bridge the shifting between shadow and non-shadow regions.

B. The MFRA Module

With the extracted convolutional perception (*i.e.*, $\mathbf{X}^j_{\text{in}} * \mathbf{W}^j$) and the shifting $\mathbf{B}^j$, how to fuse them becomes another critical question. Normally, as the network deepens, the shifting features obtained from the non-shadow regions are gradually strengthened, but the extent of this shifting attention is different at different stages in the encoder, *e.g.*, from shallow to deep stages, the shifting gradually spreads from the shadow boundary regions to the whole shadow region. The naive element-wise additive fusion ignores the feature differences between shadow and non-shadow regions in $\mathbf{B}^j$ and $\mathbf{X}^j_{\text{in}} * \mathbf{W}^j$ at different stages and fails to recover spatial feature homogeneity (see Fig. 6 $\mathbf{X}^j_{\text{in}} * \mathbf{W}^j + \mathbf{B}^j$)). Therefore, to allow each stage to account for the features of different scales fully, we adopt a multi-scale fusion strategy. Specifically, given the convolutional perception (*i.e.*, $\mathbf{X}^j_{\text{in}} * \mathbf{W}^j$) and the shifting $\mathbf{B}^j$ in Eq. 4, we first conduct multiple atrous convolutions with different dilation rates to obtain multi-scale features, *i.e.*,

$$\mathbf{X}^j_s = \sigma([\mathbf{X}^j_{\text{in}} * \mathbf{W}^j, \mathbf{B}^j] * \mathbf{D}^j_s), \tag{7}$$

where $\sigma(\cdot)$ is the ReLU function, $\mathbf{D}^j_s$ is the weight of an atrous convolution with dilation rate s. Here, we consider $s \in \mathbb{S} = \{1, 24, 12, 6\}$ and get the first three features $\{\mathbf{X}^j_s | s \in \{1, 24, 12\}\}$ via Eq. 7. For the last and smallest scale features (*i.e.*, $s = 6$), we do not extract from $\mathbf{X}^j_{\text{in}}$ like Eq. 7, but feed $\mathbf{X}^j_{12}$ to a dilation convolution to obtain $\mathbf{X}^j_6$ (see Fig. 5(d)). The size of the weights of all $\mathbf{D}^j_s$ is $3 \times 3 \times C^j_{\text{in}} \times C^j_{\text{out}}$ with strides $\{1, 1, 2, 2\}$. This implementation can alleviate heavy information loss caused by down-sampling the input features two times directly and reduce the computation cost. Then, the key problem is how to combine the four sets of features. Since different scales have different regions of interest and different stages require different scale features, we further propose dynamic weight fusion so that different stages can adaptively assign different weights to different scale features. To this end, we propose to estimate the combination parameters dynamically according to the inputs, *i.e.*,

$$\mathbf{X}^j_{\text{out}} = \sum_s^S \mathbf{w}^j_s \odot \mathbf{X}^j_s, \text{ with}$$
$$\{\mathbf{w}^j_s | s \in \mathbb{S}\} = \Phi([\mathbf{X}^j_{\text{in}} * \mathbf{W}^j, \mathbf{B}^j]), \tag{8}$$

where $\mathbf{w}^j_s \in \mathbb{R}^{H^j_{\text{in}} \times W^j_{\text{in}} \times C^j_{\text{in}}}$ assigns weights for each element in $\mathbf{X}^j_s$. Note that the elements at the same positions but different channels share the same weights. $\Phi(\cdot)$ is a subnetwork containing two convolution layers and a softmax layer. A ReLU layer follows each convolution.

Previous works, such as [31], have explored the fusion of multi-level features for shadow removal. Our multi-scale feature & residual aggregation (MFRA) approach differentiates itself from [31] in the following vital aspects: ❶ Diverse Objectives: DHAN aims primarily to learn shadows while retaining low-level details within the input image. This is achieved by aggregating multi-level features through dilated convolutions and the spatial pooling pyramid (SPP). Our MFRA module, however, is crafted to encourage homogeneity between shadow and non-shadow regions within the features. ❷ Distinct Technical Approaches: DHAN commences by extracting multi-level features and merging them, neglecting to ensure consistency between shadow and non-shadow regions at the feature level. In contrast, our MFRA is embedded within each stage of the encoder, promoting homogeneity at every step by fusing features $\mathbf{X}^j_{\text{in}} * \mathbf{W}^j$ with feature shifts $\mathbf{B}^j$. Additionally, DHAN utilizes SPP coupled with average pooling to aggregate features across different scales, a method that can inadvertently discard essential details. MFRA, on the other hand, operates at each stage without involving pooling operations in the fusion process. It also diverges from DHAN's uniform merging of multi-level features by employing dynamic fusion weights (see Eq. (8)) predicted from the

input features, allowing for a more adaptable fusion process to various inputs. A detailed comparison of our MFRA module with the SPP used by DHAN is provided in Section VI-E2.

C. Configuration Details

In this subsection, we detail the configuration of the first stage of StructNet, which contains three branches.

The first branch aims to extract the whole scene features and estimate the shadow-free structure prediction. It takes the shadow structure (*i.e.*, $\mathbf{S}_l$), the shadow mask (*i.e.*, $\mathbf{M}$), and the shifting predicted by the second branch as inputs. It consists of one encoder, one decoder, and the proposed MFRA module, where the first two parts share the same settings with the vanilla UNet in Section III-B. In terms of the fusion function (*i.e.*, MFRA), we set the kernel size of all convolutional layers to 3, and the number of kernels/filters (*i.e.*, C^j_{out}) is $\{64, 128, 256, 512, 512\}$ except for the second layer of weight generation (*i.e.*, $\Phi(\cdot)$) where the number of kernels is equal to the number of parallel branches (*i.e.*, 4).

The second branch is to estimate the shifting (*i.e.*, $\mathbf{B}^j$). It takes the shifting from the previous layer (*e.g.*, $\mathbf{B}^{j-1}$), the shadow mask from the third branch (*i.e.*, $\mathbf{M}^j_{\text{in}}$), and the jth features from the first branch (*i.e.*, $\mathbf{X}^j$) as inputs. It includes five convolutional layers, each corresponding to one of the encoder layers in the first branch and having the same strides and feature dimensions. The kernel size K^j of each layer is $\{7, 5, 3, 3, 3\}$, and each layer is followed by a Batch-Norm and a ReLU function. For the $\eta(\cdot)$, we set the kernel size and stride of the two layers as $\{K^j, 1\}$ and $\{2, 1\}$, respectively.

The third branch takes the shadow mask $\mathbf{M}$ as input and generates distinct binary masks for each layer along the encoder. We fix the constant convolutional kernels $\mathbf{W}^j_{\mathbb{1}}$ of size K^j and with stride 2.

In addition, as shown in Table I, the maximum performance gain is delivered when the structure level is 0.015. Thus, unless otherwise stated, we set $l = 0.015$ in the proposed StructNet.

V. Multi-Level StructNets (MStructNet)

Although our StructNet presented in Section IV is able to restore the shadow structure effectively and performs better than the vanilla UNet, benefiting the image-level shadow removal step significantly, such a two-stage solution leads to large computational overheads due to the naive combination of two networks. To address this problem, we further propose a self-contained shadow removal method that utilizes multi-level structures at the feature level with only a small increase in the parameter numbers. Specifically, we omit the step for predicting the shadow-free structure image through the first stage of StructNet but use the non-shadow structure information directly. We refer to this method as MStructNet, and show the pipeline in Fig. 7.

A. Pipeline

Given a shadow image $\mathbf{I}$, we extract structures via Eq. 1 and consider four levels, *i.e.*, $l \in \mathcal{L} = \{0.005, 0.015, 0.045, 0.1\}$, to obtain four levels of structure, $\{\mathbf{S}_l | l \in \mathcal{L}\}$. MStructNet takes the original shadow image, shadow mask, and all levels of structure images/layers as inputs to predict the shadow-free image directly. The whole pipeline contains three components, *i.e.*, structure-aware encoder, fusion-oriented encoder, and decoder. The structure-aware encoder contains $|\mathcal{L}|$ blocks to address $|\mathcal{L}|$ structures. Each block follows the design in StructNet to ensure that the shadow elements harmonize with the shadow-free elements. The fusion-oriented encoder consists of standard convolutions and is to further extract deep feature embedding from the structure-aware features. The decoder is to map the feature embedding to the shadow-free image. We show the whole pipeline in Fig. 7(a). As the main difference between this pipeline and the vanilla UNet lies in the structure-aware encoder as shown in Fig. 7(b), we discuss the design of this encoder in detail below.

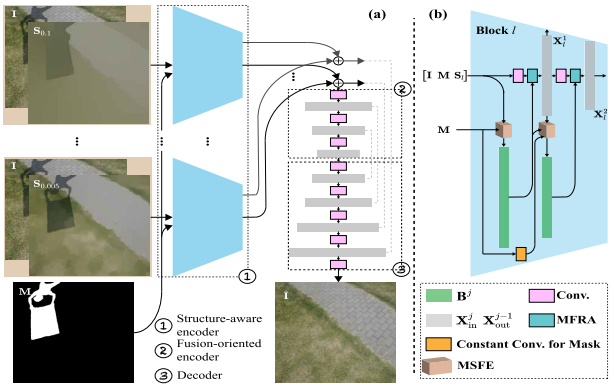


Fig. 7. Pipeline of the multi-level StructNet (MStructNet). (a) presents the whole pipeline, while (b) shows the detail of the blue blocks in (a).

In terms of the lth block in the structure-aware encoder, we have the original shadow image $\mathbf{I}$, the lth structure $\mathbf{S}_l$, and shadow mask $\mathbf{M}$ as inputs. We feed them to the block having two convolutional layers equipped with the proposed MSFE and MFRA modules (see Fig. 7(b)), which produce two features denoted as $\mathbf{X}^1_l$ and $\mathbf{X}^2_l$ corresponding to the outputs of the first and second convolution layers, respectively. For the four levels of structures (*i.e.*, $\{\mathbf{S}_l | l \in \mathcal{L}\}$), we obtain eight output features, *i.e.*, $\{\mathbf{X}^1_l\}_{l\in\mathcal{L}}$ and $\{\mathbf{X}^2_l\}_{l\in\mathcal{L}}$. We combine the four sets of features in $\{\mathbf{X}^1_l\}_{l\in\mathcal{L}}$ or $\{\mathbf{X}^2_l\}_{l\in\mathcal{L}}$ via an addition. The combined features are fed to the fusion-oriented encoder and decoder to estimate the shadow-free image $\hat{\mathbf{I}}$.

B. Configuration Details

Same as the first stage of StructNet in Section IV, each block l in the structure-aware encoder in MStructNet also has three branches. Take the lth block as an example, the inputs to the first branch include shadow image $\mathbf{I}$, shadow structure $\mathbf{S}_l$ with level l and shadow mask $\mathbf{M}$. The three inputs are concatenated along the channel axis and further fed to the standard convolution to perceive the global scene. The inputs to the second branch are shadow structure $\mathbf{S}_l$ and shadow mask $\mathbf{M}$. Then, with the global perceptual features of the first branch as the guiding weights, the shifting features can be obtained by Eq. 6. The third branch updates the shadow mask $\mathbf{M}^j$ in the same way as in Section IV-C. Regarding the fusion-oriented encoder and the decoder, they contain only standard convolutional layers, instance-norm and activation function (*e.g.*, Leaky-ReLU or ReLU), and all the settings are the same as those in the vanilla UNet in Section III-B.

VI. Experiment

A. Loss Functions

We train StructNet and MStructNet using L_1 and the perceptual loss. Given a restored image $\hat{\mathbf{I}}$ and its ground truth $\mathbf{I}^*$, we have:

$$L(\hat{\mathbf{I}}, \mathbf{I}^*) = \lambda_1 L_1(\hat{\mathbf{I}}, \mathbf{I}^*) + \lambda_2 L_{\text{perc}}(\hat{\mathbf{I}}, \mathbf{I}^*), \tag{9}$$

where the λ_1 and λ_2 are the coefficients. The $L_1(\hat{\mathbf{I}}, \mathbf{I}^*)$ is the primary loss item to supervise the training process, and we set the $\lambda_1 = 1$ by default. Regarding the $L_{\text{perc}}(\hat{\mathbf{I}}, \mathbf{I}^*)$ loss, we follow Zhang et al. [54] to set the $\lambda_2 = 0.1$. $L_1(\hat{\mathbf{I}}, \mathbf{I}^*)$ is the L_1-norm distance to ensure pixel-level visual consistency. $\mathcal{L}_{\text{perc}}$ is the perceptual loss [55], which aims to ensure the restored image has the same perception as the ground truth:

$$L_{\text{perc}}(\hat{\mathbf{I}}, \mathbf{I}^*) = \sum_{i=1}^{3} \|\text{VGG16}_i(\hat{\mathbf{I}}) - \text{VGG16}_i(\mathbf{I}^*)\|_1, \tag{10}$$

where $\text{VGG16}_i(\cdot)$ represents the activation map of the ith max-pooling layer in the VGG16 [56] pretrained on ImageNet [57].

We employ $L(\hat{\mathbf{I}}, \mathbf{I}^*)$ to end-to-end train MStructNet directly. For StructNet, we use the same loss but with $< \hat{\mathbf{S}}_l, \mathbf{S}_l^* >$ to train the first stage, *i.e.*, $L(\hat{\mathbf{S}}_l, \mathbf{S}_l^*)$. After that, we fix the parameters of the first-stage network and use $L(\hat{\mathbf{I}}, \mathbf{I}^*)$ to train the second-stage network.

B. Datasets and Metrics

1) Datasets: We conduct our experiments on three shadow removal benchmark datasets, *i.e.*, SRD [9], ISTD [11] and ISTD+ [38], to evaluate the effectiveness of the proposed methods. SRD [9] is the first large-scale shadow removal dataset, consisting of 3,088 paired shadow and shadow-free images, of which 2,680 are for training and 408 for testing. Since shadow masks are not available in SRD, we follow AEF [8] to utilize Otsu's algorithm to extract the shadow masks from the difference between the shadow and shadow-free images. We adopt the extracted masks for training and testing and use the available masks from DHAN [31] for metric evaluation. The ISTD dataset [11] contains 1,870 triplets (*i.e.*, shadow image, shadow mask, and shadow-free image) for shadow removal, with 1,330 for training and 540 for testing. Le et al. [37] later corrected the color consistency in ISTD to form the ISTD+ dataset. For both ISTD and ISTD+, we follow AEF [8] to use the ground-truth shadow masks for training and extracted masks from Otsu's algorithm for testing.

2) Evaluation Metrics: We follow methods [8], [10] to compute the root mean square error (RMSE) between the shadow-removed image and ground-truth shadow-free image in the LAB color space, which is also named image-level RMSE. When evaluating structure-level shadow removal, we compute RMSE between the predicted and ground truth structures as described in Section III-B, which is denoted as structure-level RMSE. We also report the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). In addition, we also adopt Learned Perceptual Image Patch Similarity (LPIPS) [54] to evaluate the perceptual quality of the shadow-free prediction. The lower the LPIPS, the higher the perceived quality. Note that all metrics are computed in the shadow region (S.), non-shadow regions (N. S.), and the whole image (All), respectively.

C. Method Settings

1) Baseline Methods Enhanced by StructNet: Our StructNet proposed in Section IV is able to enhance existing shadow removal methods by first conducting structure-level shadow removal, and then using the restored shadow-free structure as an auxiliary prior for the baseline method to predict the shadow-free image in the second stage. We regard four baseline methods (*i.e.*, the vanilla UNet in Section III, STC-GAN [11], AEF [8] and SADC [58]) as the second-stage networks in StructNet, resulting in four variants. We chose these methods due to their distinct frameworks, highlighting StructNet's exceptional extensibility and flexibility. Note that the four versions share the same structure-level shadow removal network. We fix the first-stage network and retrain only the second-stage networks.

To further validate the advantages of the structure-informed shadow removal networks, we compare the StructNet variants and MStructNet with two traditional methods: Guo et al. [6], Gong et al. [7], and eighteen deep learning-based methods: DeshadwoNet [9], STCGAN [11], DSC [10], MS-GAN [40], AR-GAN [59], SP+M-Net [37], CLA [32], RIS [33], Param+M+D-Net [12], DHAN [31], G2R [41], AEF [8], DC-GAN [42], SP+M+I-Net [38], BMNet [34], SADC [58], EMD-Net [39], and SGNet [35].

D. Comparisons to the State-of-the-Arts

1) Validation Results: We evaluate four StructNet variants (*i.e.*, StructNet-UNet/-STCGAN/-AEF/-SADC) on ISTD+ and ISTD by comparing them with their original versions and show the results in Table II. We can see that the proposed StructNet improves all four baselines with a significant margin on RSME in the shadow regions over the two datasets. In particular, the RMSE of STCGAN decreases from 9.39 to 6.25 (an improvement of 33.4%) on ISTD+ and from 10.11 to 7.52 (an improvement of 25.6%) on ISTD. The other three structure-enhanced methods also show clear performance boosts. As StructNet-UNet obtains the best results across all counterparts, for convenience, we refer to it as StructNet in all experiments.

2) Comparisons on Benchmarks: We compare StructNet (*i.e.*, StructNet-UNet) and MStructNet with state-of-the-art methods in three benchamrks, *i.e.*, ISTD, ISTD+ and SRD, and the results are shown in Table III, Table IV, and Table VI. Obviously, the results demonstrate that the proposed StructNet and MStructNet outperform all baseline methods in the shadow regions, showing the advantages of our structure-informed approach. Notably, StructNet obtains 6.33 in the shadow regions on the ISTD dataset, with an improvement of 16.7% and 21.6% over the BMNet and EMNet. In terms of the perceptual metric, LPIPS, StructNet also outperforms existing methods by a large margin in both ISTD and ISTD+ datasets. Meanwhile, our MStructNet achieves the lowest RMSE in shadow regions and the lowest LPIPS among competing shadow removal methods in the ISTD, ISTD+, as shown in Table III, Table IV. In particular, MStructNet has 13.1% lower

TABLE II

VALIDATION RESULTS OF STRUCTNET-EQUIPPED SHADOW REMOVAL METHODS ON ISTD AND ISTD+ DATASETS. WE EMBED FOUR EXISTING MODELS, *i.e.*, VANILLA UNET, STCGAN [11], AEF [8] AND SADC [58], IN OUR STRUCTNET FRAMEWORK AS FOUR VARIANTS, AND COMPARE THEM WITH THE ORIGINAL METHODS

Datasets	Methods	RMSE ↓			PSNR ↑			SSIM ↑		
		S.	N.S.	All	S.	N.S.	All	S.	N.S.	All
ISTD+ [37]	vanilla UNet	5.89	2.49	3.05	37.86	38.15	34.34	0.990	0.985	0.970
	StructNet-UNet	5.31	2.52	2.97	38.43	37.33	34.31	0.990	0.979	0.963
	STCGAN [11]	9.39	4.25	5.09	35.09	33.92	30.36	0.983	0.961	0.937
	StructNet-STCGAN	6.25	3.58	4.02	37.44	34.52	32.00	0.988	0.968	0.949
	AEF [8]	6.55	3.77	4.23	36.04	31.16	29.44	0.978	0.892	0.861
	StructNet-AEF	6.35	3.75	4.17	36.08	31.18	29.52	0.978	0.892	0.861
	SADC [58]	6.21	3.05	3.57	37.18	37.69	33.88	0.991	0.982	0.968
	StructNet-SADC	5.82	2.83	3.32	37.92	37.72	34.26	0.991	0.983	0.969
ISTD [11]	vanilla UNet	7.29	4.73	5.09	35.69	31.70	29.74	0.987	0.970	0.951
	StructNet-UNet	6.33	4.71	4.98	36.60	31.57	29.94	0.988	0.970	0.952
	STCGAN [11]	10.11	5.76	6.47	33.93	30.18	27.90	0.981	0.959	0.932
	StructNet-STCGAN	7.52	5.64	5.95	35.46	30.52	28.75	0.985	0.961	0.939
	AEF [8]	7.98	5.54	5.94	34.39	28.61	27.11	0.974	0.880	0.844
	StructNet-AEF	7.49	5.67	5.97	34.72	28.09	26.86	0.975	0.880	0.844
	SADC [58]	7.19	5.06	5.41	35.52	31.97	29.85	0.989	0.976	0.961
	StructNet-SADC	6.83	4.69	5.04	36.40	32.27	30.32	0.989	0.978	0.963

TABLE III

QUANTITATIVE COMPARISON WITH THE SOTA METHODS ON THE ISTD DATASET. '–' INDICATES VALUES THAT ARE NOT AVAILABLE. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Methods	RMSE ↓			PSNR ↑			SSIM ↑			LPIPS ↓
	S.	N.S.	All	S.	N.S.	All	S.	N.S.	All	
Guo *et al.* [6]	18.65	7.76	9.26	27.76	26.44	23.08	0.964	0.975	0.919	-
STCGAN [11]	10.11	5.76	6.47	33.93	30.18	27.90	0.981	0.959	0.932	0.092
MS-GAN [40]	10.57	5.91	6.67	31.73	29.02	26.36	0.980	0.959	0.928	-
DSC [10]	8.45	5.03	5.59	34.64	31.26	29.00	0.984	0.969	0.944	0.223
DHAN [31]	7.49	5.30	5.66	35.53	31.05	29.11	0.988	0.971	0.954	0.089
AR-GAN [59]	7.21	5.83	6.68	-	-	-	-	-	-	-
RIS [33]	8.99	6.33	6.95	-	-	-	-	-	-	-
CLA [32]	9.01	6.25	6.62	-	-	-	-	-	-	-
CANet [60]	8.86	6.07	6.15	-	-	-	-	-	-	-
DC-GAN [42]	10.55	5.79	6.57	31.69	28.99	26.38	0.976	0.958	0.922	0.121
BMNet [34]	7.60	4.59	5.02	35.61	**32.80**	30.28	0.988	**0.976**	**0.959**	0.089
EMNet [39]	8.08	4.75	5.22	36.27	31.85	29.98	0.986	0.965	0.944	0.087
StructNet	**6.33**	4.71	4.98	36.60	31.57	29.94	0.988	0.970	0.952	0.072
MStructNet	6.34	**4.35**	**4.68**	**36.85**	32.49	**30.65**	**0.989**	0.972	0.955	**0.059**

TABLE IV

QUANTITATIVE COMPARISON WITH THE SOTA METHODS ON THE ISTD+ DATASET. '–' INDICATES VALUES THAT ARE NOT AVAILABLE. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Method \ RMSE ↓	S.	N.S.	All	LPIPS ↓
Guo *et al.* [6]	22.0	3.1	6.1	-
Gong *et al.* [7]	13.3	-	-	-
SP+M-Net [37]	7.9	3.1	3.9	-
Param+M+D-Net [12]	9.7	3.0	4.0	0.098
G2R [41]	7.3	2.9	3.6	0.092
DC-GAN [42]	10.3	3.5	4.6	0.111
SP+M+I-Net [38]	6.0	3.1	3.6	0.080
BMNet [34]	6.1	2.9	3.5	0.079
SGNet [35]	5.9	2.9	3.4	0.091
StructNet	**5.3**	**2.5**	**3.0**	0.065
MStructNet	**5.3**	2.7	3.1	**0.054**

RMSE in the shadow region and 31.6% lower LPIPS on the ISTD+ dataset compared to BMNet. In comparison to our StructNet, MStructNet achieves better performance, with a lower RMSE in the non-shadow regions (4.35 vs. 4.71) and similar results in the shadow regions (6.34 vs. 6.33), resulting in a better overall performance (4.68 vs. 4.98) in "All" on the ISTD dataset. On the SRD dataset (see Table VI), our StructNet and MStructNet obtain the lowest RMSE results in the shadow region, and the best LPIPS perceptual quality assessments.

a) Efficiency comparisons: In Table V, we elucidate the efficiency comparisons, encompassing model parameters, floating-point operations (FLOPs), and inference time. StructNet, distinguished by its two-stage shadow removal process, exhibits slightly more parameters relative to other methods, such as DHAN [31] and G2R [41]. Although BMNet [34] operates with fewer parameters, its high-resolution processing substantially elongates the computational time, rendering it nearly ten times slower than our method. In a striking contrast, StructNet surpasses SP+M+I-Net [38], being 20 times swifter and requiring only one-fourth the FLOPs. Emphasizing the nuanced design of the first-stage network, StructNet intricately embeds MSFE and MFRA modules within each encoder layer. This leads to the increasing complexity and results in 6.32G more FLOPs than SGNet [35]. Our MStructNet improves StructNet by directly processing shadow structure at the feature level, extracting and utilizing shadow-free structure information, thereby reducing the computational overhead.

We also display the visual comparison in Fig. 8. The proposed MStructNet can effectively complement low-level cues by integrating multi-level shadow-free structure features, thus facilitating the maximum restoration of the original colors in the umbra and penumbra regions. In contrast, other methods either fail to restore the original colors (*e.g.*, MS-GAN and DC-SGAN) or cause obvious artifacts around the penumbra (*e.g.*, G2R and Param+M+D-Net).

E. Evaluation of StructNet

1) Effectiveness of the MSFE Module: We construct different StructNet variants by using different structure-level shadow removal networks and then evaluate the quality of the restored structures (*i.e.*, structure-level RMSEs) from the first stage as

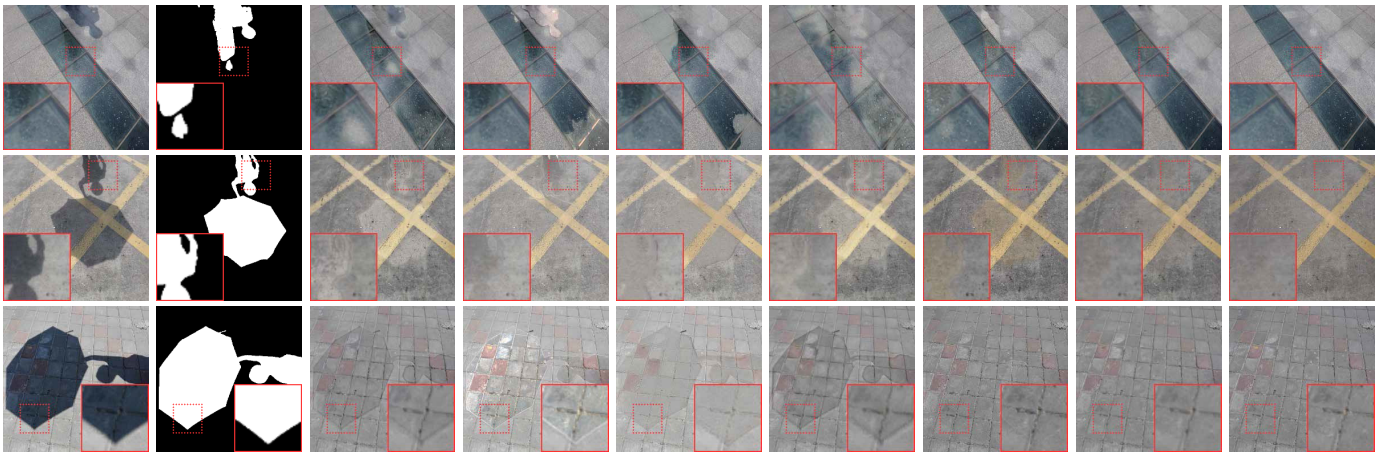


Fig. 8. Qualitative comparison on the ISTD test set. Please zoom in to see the details. Refer to the Supplemental for more visual comparisons.

TABLE V
COMPARISONS OF PARAMETERS, FLOPS, AND NETWORK INFERENCE TIME

Methods	DHAN [31]	G2R [41]	AEF [8]	DC-GAN [42]	SP+M+I-Net [37]	BMNet [34]	SGNet [35]	**StructNet**	**MStructNet**
Params. (MB)	21.75	22.76	143.01	21.16	141.18	0.37	6.17	67.06	20.62
FLOPs (G)	262.87	113.87	160.32	105.00	160.10	10.99	39.63	45.95	28.77
Time (ms)	41	59	23	6	60	33	27	3.3	2.8

TABLE VI
QUANTITATIVE COMPARISON WITH THE SOTA METHODS ON THE SRD DATASET. '–' INDICATES VALUES THAT ARE NOT AVAILABLE. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Method \ RMSE ↓	S.	N.S.	All	LPIPS ↓
Guo *et al.* [6]	29.89	6.47	12.60	-
DeShadowNet [9]	11.78	4.84	6.64	-
DSC [10]	10.89	4.99	6.23	0.248
MS-GAN [40]	-	-	7.32	-
AR-GAN [59]	7.24	4.71	5.74	-
DHAN [31]	8.39	4.67	5.46	0.197
RIS [33]	8.22	6.05	6.78	-
CLA [32]	8.10	6.01	6.59	-
DC-GAN [42]	7.70	3.39	4.66	0.109
CANet [60]	7.82	5.88	5.98	-
BMNet [34]	6.96	**3.13**	**4.18**	0.099
EMNet [39]	7.44	3.74	4.79	0.285
StructNet	6.93	3.94	4.81	0.092
MStructNet	**6.69**	4.28	4.97	**0.091**

TABLE VII
COMPARISON BETWEEN STRUCTNET VARIANTS. THE COMPARISONS ARE CONDUCTED ON THE ISTD+ DATASET FROM TWO ASPECTS, *i.e.*, STRUCTURE-LEVEL AND IMAGE-LEVEL SHADOW REMOVAL. WE DENOTE ALL VARIANTS WITH STRUCTNET(FACTOR1, FACTOR2, FACTOR3) WHERE 'FACTOR1' REPRESENTS THE FUNCTION FOR THE BRIDGE(·), 'FACTOR2' MEANS THE POSITIONS TO EMBED THE 'FACTOR1', AND 'FACTOR3' IS THE FUNCTION FOR THE FUSION(·) IN EQ. 4. "ADD" AND "CONV" REFER TO THE ADDITIVE FUSION AND CONVOLUTION OPERATIONS

Methods for the first stage \ RMSE ↓	Structure-level S.	Structure-level N.S.	Structure-level All	Image-level S.	Image-level N.S.	Image-level All
Two-stage shadow removal in Table I with $l = 0.015$	5.54	1.86	2.46	5.89	2.49	3.05
StructNet(MSFE, 1, Add)	5.10	1.87	2.41	5.72	2.59	3.10
StructNet(MSFE, 2, Add)	4.80	1.88	2.36	5.55	2.59	3.07
StructNet(MSFE, 3, Add)	4.76	2.01	2.46	5.56	2.58	3.07
StructNet(MSFE, 4, Add)	4.73	2.04	2.49	5.62	2.61	3.11
StructNet(MSFE, 5, Add)	4.74	1.97	2.43	5.55	2.57	3.06
StructNet(MSFE, 1, MFRA)	4.82	1.88	2.36	5.68	2.59	3.09
StructNet(MSFE, 2, MFRA)	4.32	1.92	2.31	5.39	2.59	3.05
StructNet(MSFE, 3, MFRA)	4.55	2.01	2.43	5.54	2.58	3.07
StructNet(MSFE, 4, MFRA)	4.65	1.88	2.33	5.57	2.52	3.02
StructNet(MSFE, 5, MFRA)	4.58	1.92	2.35	5.43	2.52	3.00
StructNet(MSFE, $(1 \cdots 5)$, Add)	4.82	1.87	2.35	5.50	2.60	3.07
StructNet(MSFE, $(1 \cdots 5)$, MFRA)	4.20	1.71	2.12	5.31	2.52	2.97

well as the quality of the restored images (*i.e.*, image-level RMSEs) from the second stage.

a) Adding MSFE to one single convolution layer: To avoid the influence of the fusion function carried out by the MFRA module, we replace it with a naive element-wise additive operation instead. We use StructNet(MSFE, j, Add) to denote the StructNet whose first-stage network uses the MSFE as Bridge() at the jth layer and the element-wise additive operation as Fusion(). We then obtain five variants, *i.e.*, {StructNet(MSFE, j, Add)$|\, j \in \{1, 2, 3, 4, 5\}$}. Table VII shows the results and we observe: ❶ Compared with the naive two-stage shadow removal method (*i.e.*, vanilla UNet), StructNets with a single MSFE achieves lower structure-level and image-level RMSEs (*i.e.*, StructNet(MSFE, 1/2/3/4/5, Add) in Table VII vs. two-stage shadow removal in Table I) in the shadow regions, which demonstrates that the MSFE does benefit the structure-level shadow removal and enhance the image-level shadow removal. ❷ In general, if we embed MSFE in a deeper convolution layer, we get lower RMSEs in the shadow regions while slightly higher RMSEs in the non-shadow regions at the structure level. For example, the

TABLE VIII
COMPARING STRUCTNET VARIANTS WITH MSFE, PARTIAL CONVOLUTION, AND GATED CONVOLUTION

Methods for the first stage \ RMSE ↓	Structure-level S.	Structure-level N.S.	Structure-level All	Image-level S.	Image-level N.S.	Image-level All
Two-stage shadow removal in Table I with $l = 0.015$	5.54	1.86	2.46	5.89	2.49	3.05
StructNet(CONVSkip)	5.10	1.97	2.48	5.82	2.61	3.14
StructNet(PartialCONV)	7.99	1.99	2.89	6.50	2.57	3.22
StructNet(GatedCONV)	5.46	1.98	2.55	6.27	2.55	3.16
StructNet(MSFE)	4.20	1.71	2.12	5.31	2.52	2.97

structure-level RMSE of the shadow region decreases from 5.10 to 4.73 if we add MSFE from the 1st to the 5th layers. We have similar observations on the image-level RMSEs.

b) Adding MSFE to all convolution layers: We further add MSFE to all layers and denote this variant as StructNet(MSFE, $(1 \cdots 5)$, Add). Compared with StructNet(MSFE, 5, Add), StructNet(MSFE, $(1 \cdots 5)$, Add) has a lower structure-level RMSE (*i.e.*, 1.87) in the non-shadow regions but a slightly higher structure-level RMSE (*i.e.*, 4.82) in the shadow regions. The overall RMSE becomes 2.35, which is smaller than that of StructNet(MSFE, 5, Add). In contrast, compared with StructNet(MSFE, 1, Add), StructNet(MSFE, $(1 \cdots 5)$, Add) has a much lower structure-level RMSE in the shadow regions and the same RSME in the non-shadow regions. Such observations imply that equipping more convolutions with MSFE can balance the restoration in the shadow and non-shadow regions.

c) Comparison with other representative convolutions: We extend our comparison of the MSFE to include three prominent convolutional structures: convolution with a skip function, partial convolution [61], and gated convolution [62]. ❶ We consider the convolution with a skip function and implement a variant denoted as StructNet(CONVSkip). In this variant, the fusion function is formulated as an additive operation, and the bridge is designed as a convolutional layer, akin to the convolutional skip connection used in residual networks [63]. ❷ In terms of the partial convolution, we follow the way in [61] and reformulate Eq. (4) and Eq. (5) as $\mathbf{X}^j_{\text{out}} = \mathbf{B}^j$ and $\mathbf{B}^j = \text{Bridge}(\mathbf{X}_{\text{in}}, \mathbf{M}^j_{\text{in}}) = \alpha_{\mathbf{p}} \sum_{\mathbf{q} \in \mathcal{N}_{\mathbf{p}}} \mathbf{X}_{\text{in}}[\mathbf{q}](1 - \mathbf{M}^j_{\text{in}}[\mathbf{q}])\mathbf{W}'^j_{\text{B}}[\mathbf{q} - \mathbf{p}]$, respectively. We denote the method as StructNet(PartialCONV). ❸ In terms of the gated convolution, we follow [62] and reformulate Eq. (4) and Eq. (5) as $\mathbf{X}^j_{\text{out}} = \mathbf{B}^j \odot (\mathbf{X}^j_{\text{in}} * \mathbf{W}^j)$ and $\mathbf{B}^j = \text{Sigmoid}(\mathbf{X}^j_{\text{in}} * \mathbf{W}^j_{\text{f}})$. We denote the method as StructNet(GatedCONV).

We present the comparative results in Table VIII and identify the following key observations: ❶ Our method, leveraging MSFE, surpasses all three baseline methods within shadow and non-shadow regions at both the structural and image levels, thereby substantiating the benefits of MSFE. ❷ The variant StructNet(CONVSkip) employing a convolutional skip function yields lower RMSEs than the naive two-stage shadow removal method within shadow regions but exhibits higher RMSEs in non-shadow areas. This dichotomy illustrates its potency in shadow removal but underlines an adverse impact on non-shadow regions. This discrepancy occurs primarily because the convolution for the bridge function and the element-wise skip function for fusion inadequately address the shift between shadow and non-shadow regions, failing to overcome the constraints of standard convolution. ❸ The approach StructNet(PartialCONV) registers considerably higher RMSEs compared to the naive two-stage method at both structural and image levels. The root cause of this deterioration is the total disregard of the original information within the shadow regions, culminating in marked performance degradation. ❹ Lastly, StructNet(GatedCONV) achieves lower RMSEs than the naive two-stage method at the structural level, but incurs higher RMSEs at the image level. This pattern corroborates the capability of gated convolution to restore structural information effectively, while also highlighting its failure to recover finer details.

d) Feature comparison: We also compare the proposed MSFE with the standard convolution in Fig. 4 by showing their processed features (See Fig. 4(e) vs. (c)). Clearly, the visual feature differences between shadow and non-shadow regions of our StructNet are much smaller than those of the naive UNet. In addition, as depicted in Fig. 4(f), the proposed StructNet presents much smaller absolute feature differences, which also demonstrates its effectiveness.

2) Effectiveness of the MFRA Module:

a) Adding MFRA to the convolution in the MSFE-based network: We replace the element-wise additive fusion (*i.e.*, "Add") of the variants in Table VII (*i.e.*, StructNet(MSFE, $\star$, Add)) with the proposed MFRA to obtain new variants, StructNet(MSFE, $\star$, MFRA), where '$\star$' denotes specific layer indexes used by StructNet. We have the following observations: ❶ All single-MSFE-based variants with MFRA (*i.e.*, StructNet(MSFE, $\star$, MFRA)) outperform the variants with the element-wise addition operation, which demonstrates that the proposed aggregation function does enhance shadow removal significantly. For example, StructNet(MSFE, 2, MFRA) achieves 4.32 structure-level RMSE in the shadow regions, outperforming StructNet(MSFE, 2, Add) by 10.0%. ❷ When we embed MFRA to all convolutions with MSFE, we find that StructNet(MSFE, $(1 \cdots 5)$, MFRA) achieves much better restoration quality in both shadow and non-shadow regions than StructNet(MSFE, $(1 \cdots 5)$, Add).

b) Comparison with alternative fusion solutions: We further compare the proposed MFRA with three potential fusion approaches to validate its advantages and effectiveness by comparing the structure restoration quality (*i.e.*, structure-level RMSE). *First*, we substitute MFRA with the ASPP [64] and denote this variant as StructNet(MSFE, $(1 \cdots 5)$, ASPP). *Second*, we degrade MFRA by removing the dynamic weights $\mathbf{B}^j_s$ in Eq. 8 and adding different scale features directly. We denote this variant as StructNet(MSFE, $(1 \cdots 5)$, MFRA_{v1}). *Third*, we construct a degraded variant of MFRA to calculate all four scale features in Section IV-B through Eq. 7 directly, and we name it as StructNet(MSFE, $(1 \cdots 5)$, MFRA_{v2}).

We report the comparison results in Table IX and have the following conclusion: ❶ Compared with the baseline fusion strategy StructNet(MSFE, $(1 \cdots 5)$, Add), StructNet(MSFE, $(1 \cdots 5)$, ASPP) obtains a larger structure-level RMSE in the shadow regions, which implies that naively using ASPP is not good enough to fuse multi-scale features for shadow removal. ❷ Compared with

TABLE IX

ABLATION STUDY ON THE PROPOSED MFRA MODULE. STRUCTNET (MSFE, $(1\cdots5)$, MFRA_{V1}) IS THE DEGRADED MFRA BY REMOVING THE DYNAMIC WEIGHTS $\mathbf{B}_s^j$ IN EQ. 8 AND ADDING DIFFERENT SCALE FEATURES DIRECTLY. WE INCLUDE ANOTHER DEGRADED VARIANT OF MFRA (*i.e.*, STRUCTNET(MSFE, $(1\cdots5)$, MFRA_{V2})) BY COMPUTING ALL FOUR SCALE FEATURES THROUGH EQ. 7 DIRECTLY

Variants \ Structure-level RMSE ↓	S.	N.S.	All
StructNet(MSFE, $(1\cdots5)$, Add)	4.82	1.87	2.35
StructNet(MSFE, $(1\cdots5)$, ASPP)	5.15	1.81	2.37
StructNet(MSFE, $(1\cdots5)$, MFRA_{v1})	4.65	1.79	2.26
StructNet(MSFE, $(1\cdots5)$, MFRA_{v2})	4.42	1.82	2.25
StructNet(MSFE, $(1\cdots5)$, MFRA)	4.20	1.71	2.12

TABLE X

ABLATION EXPERIMENT OF MSTRUCTNET ON THE ISTD+ DATASET, WITH RESPECT TO DIFFERENT STRUCTURE LEVEL UTILIZATION

Structure levels				Image-level RMSE ↓		
0.005	0.015	0.045	0.1	S.	N.S.	All
✓				5.46	2.90	3.32
	✓			5.46	2.78	3.22
		✓		5.61	2.73	3.20
			✓	5.66	2.74	3.21
✓	✓			5.46	2.81	3.24
✓	✓	✓		5.42	2.79	3.22
✓	✓	✓	✓	5.29	2.73	3.15

the degraded version StructNet(MSFE, $(1\cdots5)$, MFRA_{v1}), StructNet(MSFE, $(1\cdots5)$, MFRA) obtains lower RMSEs in both shadow and non-shadow regions, leading to a lower RMSE in "All" (*i.e.*, 2.12 vs. 2.26), which demonstrates that combining multi-scale features with dynamically predictive weights via Eq. 8 indeed helps restore the structure better. ❸ Using our proposed strategy for extracting the smallest scale features prevents heavy information loss during down-sampling, as shown by the lower RMSEs in shadow and non-shadow regions (4.20 to 4.42 and 1.71 to 1.82, respectively).

F. Effectiveness of MStructNet

1) Numbers of Structure Levels: In Section V-A, the structure-aware encoder is made up of several blocks, with each block representing one structure level and containing two convolution layers equipped with the MSFE and MFRA modules. Note that we set two convolution layers for each block due to the empirical results in Table VII (as StructNet(MSFE, 2, Add/MFRA) achieves the lowest RMSE in "All", among single-convolution based variants).

2) Number of Blocks (or Structure Levels) in MStructNet: As discussed in Section V-A, each block contains MSFE and MFRA modules to form a structure level, and the final MStructNet fuses structures of all different levels. Here, we study the effects of using different numbers of blocks in the structure-aware encoder to validate the advantages of exploiting the multi-level structures. Specifically, we may obtain four variants of MStructNet by using a single structure selected from $\{0.005, 0.015, 0.045, 0.1\}$, and are denoted as: $\{\text{MStructNet}(l)|l \in \{0.005, 0.015, 0.045, 0.1\}\}$. We then gradually add more structures to MStructNet(0.005) to obtain three more variants, denoted as: MStructNet($\{0.005, 0.015\}$), MStructNet($\{0.005, 0.015, 0.045\}$), and MStructNet($\{0.005, 0.015, 0.045, 0.1\}$), respectively. The last version denotes the final version of MStructNet. As reported in Table X, we can see that: ❶ MStructNet with the structure level 0.015 shows the best results among all single structure level variants. ❷ when we add more structure levels, the restoration quality gradually improves and MStructNet with all four structure levels achieves the lowest RMSE in the shadow and non-shadow regions, which confirms that the utilization of multi-level non-shadow structures at the feature level can indeed benefit the image-level shadow removal.

TABLE XI

COMPARISONS OF DIFFERENT LOSS RATIOS FOR PERCEPTUAL LOSS. THE $\lambda_1 = 1$ BY DEFAULT. THE MSTRUCTNET MODEL AND THE ISTD+ DATASET WERE UTILIZED FOR THESE EXPERIMENTS

λ_2 ratio	RMSE ↓			PSNR ↑			SSIM ↑			LPIPS ↓
	S.	N.S.	All	S.	N.S.	All	S.	N.S.	All	
0.01	5.45	2.66	3.12	38.05	36.00	33.39	0.990	0.976	0.960	0.062
0.1 (Ours)	**5.34**	2.73	**3.10**	**38.27**	35.83	33.34	**0.990**	0.976	0.962	**0.054**
1	5.55	2.73	3.19	38.00	36.46	33.57	0.990	0.977	0.963	0.059
10	5.64	2.79	3.26	38.03	36.61	33.68	0.990	0.977	0.962	0.059

3) Different Loss Ratio L_{perc} in Eq. (9): We have also conducted experiments using different combinations of loss ratios, *i.e.*, $\lambda_2 \in \{0.01, 0.1, 1, 10\}$, and the results are presented in Table XI. It is worth noting that the ratio combination of $\lambda_1 = 1$ and $\lambda_2 = 0.1$ achieves the best results in all metrics in the shadow regions. Note that we have not conducted meticulous adjustments to the loss ratios.

VII. CONCLUSION

In this paper, we have systematically investigated the utilization and efficacy of image structure for single-image shadow removal. *First*, we have built vanilla UNet-based networks to restore the shadow-free structure of the input shadow image, and revealed that image structure can help enhance the quality of shadow-removed images significantly. *Second*, we have proposed a novel two-stage removal network named structure-informed shadow removal network (StructNet). It includes two new modules for the utilization of structure information, *i.e.*, *mask-guided shadow-free extraction (MSFE) module* and *multi-scale feature & residual aggregation (MFRA) module*, to extract the image structural features and regularize the feature consistency, respectively. We have shown that StructNet can help improve the performances of three state-of-the-art methods. *Third*, based on StructNet, we have further proposed a self-contained shadow removal method to fully excavate the potential of multi-level structures at the feature level, named *multi-level StructNets (MStructNet)*, which has fewer parameters and low computational costs. The extensive results on three public datasets have also demonstrated the advantages and effectiveness of the proposed StructNet and MStructNet.

REFERENCES

[1] C. Bouganis and M. Brookes, "Multiple light source detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 4, pp. 509–514, Apr. 2004.

[2] W. Zhang, X. Zhao, J.-M. Morvan, and L. Chen, "Improving shadow suppression for illumination robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 611–624, Mar. 2019.

[3] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *Proc. CVPR*, 2020, pp. 2777–2787.

[4] *Adobe Photoshop CS6*, Wiley, Hoboken, NJ, USA, 2012.

[5] M. Gryka, M. Terry, and G. J. Brostow, "Learning to remove soft shadows," *ACM Trans. Graph.*, vol. 34, no. 5, pp. 1–15, Nov. 2015.

[6] 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.

[7] H. Gong and D. Cosker, "Interactive removal and ground truth for difficult shadow scenes," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 33, no. 9, p. 1798, 2016.

[8] L. Fu et al., "Auto-exposure fusion for single-image shadow removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10571–10580.

[9] 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.

[10] 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.

[11] 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.

[12] H. Le and D. Samaras, "From shadow segmentation to shadow removal," in *Proc. ECCV*, 2020, pp. 264–281.

[13] R. Arnheim, *Art and Visual Perception: A Psychology of the Creative Eye*. Berkeley, CA, USA: Univ of California Press, 1954.

[14] J. Johnson, "Our vision is optimized to see structure," DMM, Tech. Rep., 2010.

[15] L. Karacan, E. Erdem, and A. Erdem, "Structure-preserving image smoothing via region covariances," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 1–11, Nov. 2013.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[17] M. S. Drew, G. D. Finlayson, and S. D. Hordley, "Recovery of chromaticity image free from shadows via illumination invariance," in *Proc. ICCVW*, 2003, pp. 32–39.

[18] G. D. Finlayson and M. S. Drew, "4-sensor camera calibration for image representation invariant to shading, shadows, lighting, and specularities," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, Jul. 2001.

[19] 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.

[20] G. D. Finlayson, S. D. Hordley, and M. S. Drew, "Removing shadows from images using retinex," in *Proc. CIC*, 2002, pp. 73–79.

[21] 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.

[22] 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.

[23] M. Xu, J. Zhu, P. Lv, B. Zhou, M. F. Tappen, and R. Ji, "Learning-based shadow recognition and removal from monochromatic natural images," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5811–5824, Dec. 2017.

[24] G. D. Finlayson, M. S. Drew, and C. Lu, "Entropy minimization for shadow removal," *Int. J. Comput. Vis.*, vol. 85, no. 1, pp. 35–57, Oct. 2009.

[25] 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.

[26] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[27] Y. Liu et al., "Tripartite information mining and integration for image matting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7555–7564.

[28] F. Liu et al., "Referring image segmentation using text supervision," in *Proc. ICCV*, 2023, pp. 22124–22134.

[29] Q. Guo, X. Li, F. Juefei-Xu, H. Yu, Y. Liu, and S. Wang, "JPGNet: Joint predictive filtering and generative network for image inpainting," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 386–394.

[30] X. Li, Q. Guo, D. Lin, P. Li, W. Feng, and S. Wang, "MISF: Multi-level interactive Siamese filtering for high-fidelity image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1869–1878.

[31] X. Cun, C.-M. Pun, and C. Shi, "Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN," in *Proc. AAAI*, 2020, pp. 10680–10687.

[32] L. Zhang, C. Long, Q. Yan, X. Zhang, and C. Xiao, "CLA-GAN: A context and lightness aware generative adversarial network for shadow removal," *Comput. Graph. Forum*, vol. 39, no. 7, pp. 483–494, Oct. 2020.

[33] L. Zhang, C. Long, X. Zhang, and C. Xiao, "RIS-GAN: Explore residual and illumination with generative adversarial networks for shadow removal," in *Proc. AAAI*, 2020, pp. 12829–12836.

[34] Y. Zhu, J. Huang, X. Fu, F. Zhao, Q. Sun, and Z.-J. Zha, "Bijective mapping network for shadow removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5627–5636.

[35] J. Wan, H. Yin, Z. Wu, X. Wu, Y. Liu, and S. Wang, "Style-guided shadow removal," in *Proc. ECCV*, 2022, pp. 361–378.

[36] X. Li et al., "Leveraging inpainting for single-image shadow removal," in *Proc. ICCV*, 2023, pp. 13055–13064.

[37] 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.

[38] H. Le and D. Samaras, "Physics-based shadow image decomposition for shadow removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9088–9101, Dec. 2022.

[39] Y. Zhu, Z. Xiao, Y. Fang, X. Fu, Z. Xiong, and Z.-J. Zha, "Efficient model-driven network for shadow removal," in *Proc. AAAI*, 2022, pp. 3635–3643.

[40] 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.

[41] Z. Liu, H. Yin, X. Wu, Z. Wu, Y. Mi, and S. Wang, "From shadow generation to shadow removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4927–4936.

[42] Y. Jin, A. Sharma, and R. T. Tan, "DC-ShadowNet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5027–5036.

[43] Z. Liu, H. Yin, Y. Mi, M. Pu, and S. Wang, "Shadow removal by a lightness-guided network with training on unpaired data," *IEEE Trans. Image Process.*, vol. 30, pp. 1853–1865, 2021.

[44] Y. He, Y. Xing, T. Zhang, and Q. Chen, "Unsupervised portrait shadow removal via generative priors," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 236–244.

[45] N. Inoue and T. Yamasaki, "Learning from synthetic shadows for shadow detection and removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4187–4197, Nov. 2021.

[46] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "StructureFlow: Image inpainting via structure-aware appearance flow," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 181–190.

[47] S. Gui, C. Wang, Q. Chen, and D. Tao, "FeatureFlow: Robust video interpolation via structure-to-texture generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14004–14013.

[48] X. Wang and J. Yu, "Learning to cartoonize using white-box cartoon representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8090–8099.

[49] L. Xu, Q. Yan, Y. Xia, and J. Jia, "Structure extraction from texture via relative total variation," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–10, Nov. 2012.

[50] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.

[51] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, no. 1, 2013, p. 3.

[52] A. W. Harley, K. G. Derpanis, and I. Kokkinos, "Segmentation-aware convolutional networks using local attention masks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5038–5047.

[53] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. ECCV*, 2018, pp. 85–100.
[54] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[55] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016, pp. 694–711.
[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[58] Y. Xu, M. Lin, H. Yang, F. Chao, and R. Ji, "Shadow-aware dynamic convolution for shadow removal," 2022, *arXiv:2205.04908*.
[59] 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.
[60] Z. Chen, C. Long, L. Zhang, and C. Xiao, "CANet: A context-aware network for shadow removal," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4743–4752.
[61] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. ECCV*, 2018, pp. 85–100.
[62] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4471–4480.
[63] 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.
[64] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

Yuhao Liu received the B.Eng. degree from Zhengzhou University in 2019 and the M.Sc. degree from the Dalian University of Technology in 2022. He is currently pursuing the Ph.D. degree in computer science with the City University of Hong Kong, where his research is focused on solving computer vision and image processing problems.

Qing Guo (Member, IEEE) received the Ph.D. degree in computer application technology from the School of Computer Science and Technology, Tianjin University, China. He was a Research Fellow with Nanyang Technology University, Singapore, from December 2019 to August 2020, and from December 2021 to September 2022. He was the Wallenberg-NTU Presidential Postdoctoral Fellow with Nanyang Technological University, from September 2020 to December 2021. He is currently a Research Scientist with the Center for Frontier AI Research, Agency for Science, Technology, and Research (A*STAR), Singapore. His research interests include computer vision, AI security, and image processing.

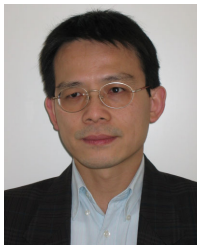
Lan Fu received the M.S. degree in biomedical engineering from Tianjin University, Tianjin, China, and the Ph.D. degree in computer science and engineering from the University of South Carolina, Columbia, SC, USA. Currently, she is a Senior Research Engineer with InnoPeak Technology Inc., Palo Alto, CA, USA. Her research interests include computer vision, deep learning, and image processing.

Zhanghan Ke (Graduate Student Member, IEEE) received the B.Eng. degree from Northeastern University, China. He is currently pursuing the Ph.D. degree with the City University of Hong Kong. His research interests include semi and self-supervised learning and its applications in computer vision. He serves as a reviewer for several computer vision conferences (e.g., CVPR, ICCV, and ECCV) and journals (e.g., IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *IJCV*, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY).

Ke Xu received the dual Ph.D. degrees from the Dalian University of Technology and the City University of Hong Kong. He is currently with the Department of Computer Science, City University of Hong Kong. His research interests include deep learning, object detection, and image enhancement and editing.

Wei Feng (Member, IEEE) received the Ph.D. degree in computer science from the City University of Hong Kong in 2008. From 2008 to 2010, he was a Research Fellow with The Chinese University of Hong Kong and the City University of Hong Kong. He is currently a Full Professor with the School of Computer Science and Technology, College of Computing and Intelligence, Tianjin University, China. His major research interests are active robotic vision and visual intelligence, specifically active camera relocalization and lighting recurrence, general Markov random fields modeling, energy minimization, active 3D scene perception, SLAM, video analysis, and generic pattern recognition. Recently, he focuses on solving preventive conservation problems of cultural heritages via computer vision and machine learning. He is an Associate Editor of *Neurocomputing* and the *Journal of Ambient Intelligence and Humanized Computing*.

Ivor W. Tsang (Fellow, IEEE) has been the Director of the Centre for Frontier AI Research (CFAR), A*STAR, since Jan 2022. Previously, he was a Professor of artificial intelligence with the University of Technology Sydney (UTS) and the Research Director of the Australian Artificial Intelligence Institute (AAII). He serves as the Senior Area Chair/the Area Chair for NeurIPS, ICML, AAAI, and IJCAI; and serves on the Steering Committee of ACML. He serves on the Editorial Board for the *Journal of Machine Learning Research*, *Machine Learning*, *Journal of Artificial Intelligence Research*, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE, IEEE TRANSACTIONS ON BIG DATA, and IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE.



Rynson W. H. Lau (Senior Member, IEEE) received the Ph.D. degree from the University of Cambridge. He was on the Faculty of Durham University. He is currently with the City University of Hong Kong. His research interests include computer graphics and computer vision. He has served on the Committee for a number of conferences, including the Program Co-Chair for ACM VRST 2004, ACM MTDL 2009, and IEEE U-Media 2010; and the Conference Co-Chair for CASA 2005, ACM VRST 2005, ACM MDI 2009, and ACM VRST 2014. He serves on the Editorial Board for the *International Journal of Computer Vision* (IJCV) and *Computer Graphics Forum*. He has served as a Guest Editor for a number of journal special issues, including *ACM Transactions on Internet Technology*, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, and IEEE COMPUTER GRAPHICS AND APPLICATIONS.