This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3252271

A Decoupled Multi-Task Network for Shadow Removal

Jiawei Liu* , Qiang Wang* , Huijie Fan ,Wentao Li, Liangqiong Qu , Yandong Tang

***Abstract*—Shadow removal, which aims to restore the illumination in shadow regions, is challenging due to the diversity of shadows in terms of location, intensity, shape, and size. Different from most multi-task methods, which design elaborate multi-branch or multi-stage structures for better shadow removal, we introduce feature decomposition to learn better feature representations. Specifically, we propose a single-stage and decoupled multi-task network (DMTN) to explicitly learn the decomposed features for shadow removal, shadow matte estimation, and shadow image reconstruction. First, we propose several coarse-to-fine semi-convolution (SMC) modules to capture features sufficient for joint learning of these three tasks. Second, we design a theoretically supported feature decoupling layer to explicitly decouple the learned features into shadow image features and shadow matte features via weight reassignment. Last, these features are converted to a target shadow-free image, affiliated shadow matte, and shadow image, supervised by multi-task joint loss functions. With multi-task collaboration, DMTN effectively recovers the illumination in shadow areas while ensuring the fidelity of non-shadow areas. Experimental results show that DMTN competes favorably with state-of-the-art multi-branch/multi-stage shadow removal methods, while maintaining the simplicity of single-stage methods. We have released our code to encourage future exploration in powerful feature representation for shadow removal (https://github.com/nachifur/DMTN).**

***Index Terms*—Shadow removal, multi-task, feature decoupling, illumination compensation, decomposition.**

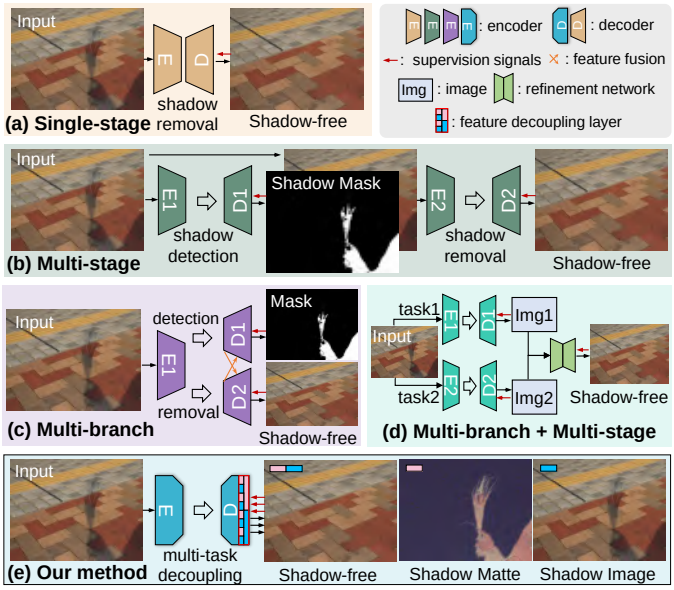

Fig. 1. Overview of the existing shadow removal networks. Our method decouples the learned features into three task domains by the proposed feature decoupling layer, while other multi-task methods learn the features corresponding to each task by redundant feature encoding and decoding.

I. Introduction

THE shadow is a ubiquitous physical phenomenon in nature and is formed when direct illumination is blocked by an object. Shadows often degrade the performance of some computer vision tasks, such as segmentation, detection, recognition, and tracking [1]–[10]. Shadow removal can be incorporated into these tasks to improve their robustness to direct illumination variations. Early approaches [11]–[17] determined illumination parameters to remove shadows by physically modeling them. These methods, however, highly rely on prior knowledge (such as illumination conditions and gradients [11], [18], [19]), and often work poorly in the umbra or penumbra regions.

Recently published large-scale datasets [20]–[25] have stimulated the emergence of two types of deep learning-based methods for shadow removal, including single-task and multi-task networks. Single-task methods [20], [26] apply a single-stage structure, such as in Fig. I (a), while multi-task learning methods usually have a multi-stage [21], [22], [27], [28], multi-branch [24], or multi-branch+multi-stage" [29]–[31] structure, as shown in Fig. I (b)–(d). A single-stage shadow removal network, while simple and efficient, may not be optimal for scenes with complex backgrounds and shadows. Consequently, recent research has favored multi-task networks that leverage complementary information from multiple tasks (e.g., shadow detection [21], [24], [27], matte estimation [22], [29], and shadow generation [23], [31]–[34]) to further improve the performance of shadow removal.

Even with variant network structures, these two types of

This work is supported by the National Natural Science Foundation of China (61991413, 62073205, U20A20200, 61873259), and the Youth Innovation Promotion Association of Chinese Academy of Sciences(2019203).

Huijie Fan is the corresponding author of this work.

*The first two authors contributed equally to this work.

J. Liu is with State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China; Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110169, China; University of Chinese Academy of Sciences, Beijing, 100049, China (email: liujiawei@sia.cn).

Q. Wang is with the Key Laboratory of Manufacturing Industrial Integrated, Shenyang University, Shenyang, 110044, China; the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China (email: wangqiang@sia.cn).

H. Fan, W. Li and Y. Tang are with State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China; Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110169, China (email: {fanhuijie,liwentao,ytang}@sia.cn).

L. Qu is with the Department of Statistics and Actuarial Science and the Institute of Data Science, The University of Hong Kong, Hong Kong, 999077 (e-mail: liangqqu@hku.hk).

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3252271

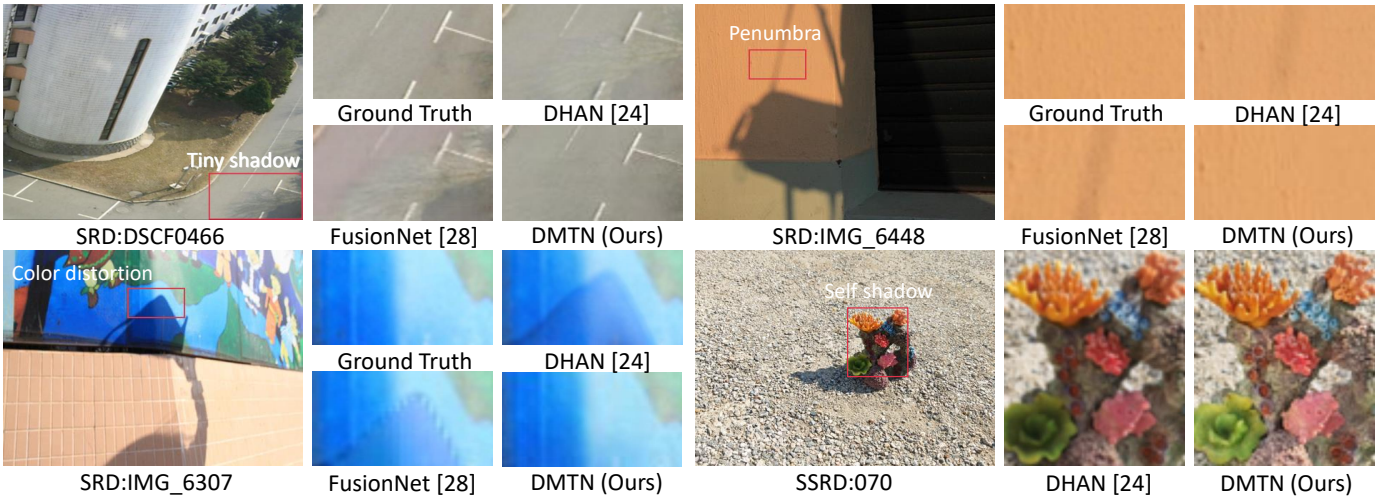

Fig. 2. Visual comparison results of removing tiny shadows, penumbras (on SRD [20]) and self-shadows (on our SSRD). With multi-task collaboration, our method better removes penumbras and tiny shadows.

methods share an underlying shadow removal principle: input a shadow image, learn the features from it, aggregate them, and output its corresponding shadow-free image. This indicates that it is not trivial to directly learn the features from the shadow image to sufficiently represent a target shadow-free image through a single-stage network. Multi-task methods can learn such features more efficiently through complex task decomposition or multi-task collaboration, resulting in superior performance. However, the widely applied multi-branch/multi-stage structures in multi-task methods may lead to redundant encoding/decoding modules or unsatisfactory performance, since they may be vulnerable to erroneous performance of the previous stage.

In this paper, instead of an elaborately designed multi-branch or multi-stage network, we introduce feature decomposition to learn better feature representations for the target shadow-free image. Motivated by the well-known image decomposition theory in the image domain,[1] we argue for a similar decomposition theory in the feature space, i.e., the features for the target shadow-free image can be decomposed as shadow image features and shadow matte[2] features. Hence we propose a single-stage, decoupled multi-task network (DMTN) to explicitly learn the decomposed features for better shadow removal, as shown in Fig. I (e) and Fig. 3.

Specifically, we first propose several coarse-to-fine semi-convolution (SMC) modules to progressively convert the original shadow features to shadow matte features, while the remaining features are untouched to represent the shadow image. Although such a structure allows us to learn the features that can represent the shadow image and shadow matte, it is still unclear which part of the features represent shadow image and which part represents the shadow matte. We then design a theoretically supported feature decoupling layer to explicitly decouple the learned features into shadow image features and shadow matte features via weight reassignment. Finally, these features are converted to the target shadow-free image, affiliated shadow matte, and shadow image, supervised by multi-task joint loss functions.

Experimental results show that our DMTN competes favorably with state-of-the-art multi-branch/multi-stage shadow removal methods (see Fig. 2), while maintaining the simplicity of a single-stage network. We envision our research could open new possibilities for the design of a better shadow removal network, highlighting the importance of good feature representation for target shadow-free images.

[1]Image decomposition: a target image can be decomposed as an input image and residual image, as derived from residual learning [35]. This is widely used in image restoration [29], [34], [36]–[42].

[2]The shadow matte here is defined as the residual between the target shadow-free image and the input shadow image.

II. Related Work

We review the development process of single image shadow removal from shadow modeling to deep learning.

Model-based Shadow Removal: Early methods [11]–[16] solved illumination parameters to remove shadows by physical modeling, based on the prior information of shadow positions. However, removing shadows with shadow detection cannot handle penumbras. Another way to cope with uneven illumination is intrinsic decomposition [43]–[46], but the pixel values of the intrinsic image are changed in the non-shadow areas.

Deep-Learning-based Shadow Removal: We first review the single-stage shadow removal network. Our previous work [20] proposed the first end-to-end network to remove shadows by estimating illumination attenuation in shadow regions, which is a multi-context architecture based on intrinsic image decomposition. Hu *et al.* [26] proposed a single-stage multi-branch network to detect or remove shadows at multiple scales by capturing directional features.

Recent research has favored multi-task shadow removal to further improve the performance of single-stage shadow removal networks. This makes sense because shadow removal aims to recover shadow regions without changing the color of non-shadow regions, which implies the importance of shadow location information and the rationality of decomposing the shadow removal task into multiple stages. Khan *et al.* [7] applied multiple CNNs to detect shadows, and a Bayesian

model to remove them. Wang *et al.* [21] proposed a stacked generative adversarial network (GAN) for joint learning of shadow detection and shadow removal in a unified manner. Ding *et al.* [27] proposed a recurrent GAN-based network consisting of multiple cascaded detection and removal networks to progressively remove shadows. Subsequently, some researchers used the shadow mask detected by the shadow detection network as prior information to improve the performance of shadow removal. Le and Samaras [22], [25] proposed a two-stage network to estimate the physical illumination parameters to remove shadows based on the shadow linear model [13] and shadow image decomposition [47]. Fu *et al.* [28] modeled shadow removal as a multiple exposure problem and proposed a three-stage network consisting of exposure estimation, exposure fusion, and boundary-aware refinement networks. Zhu *et al.* [48] proposed an interpretable unfolding network for shadow removal, based on a shadow removal optimization algorithm.

However, cascaded multi-stage networks are susceptible to the output of the previous stage. Cun *et al.* [24] proposed a single-stage network for shadow detection and removal, which can alleviate the dependency of auxiliary tasks through a parallel branch structure. Zhang *et al.* [29] proposed a shadow removal network without shadow detection by introducing more complementary information. They used three networks in parallel to estimate the residual, illumination, and coarse shadow-free images, and fused them for shadow removal by an encoder-decoder structure. Chen *et al.* [30] proposed a two-stage shadow removal network based on block matching. In the first stage, they used two parallel encoders to extract the shadow image features and obtain contextual matching pairs from the shadow unaware image, transferred the features of non-shadow regions to shadow regions, and obtained a coarse shadow-free image. The second stage used a refinement network. Zhu *et al.* [31] used shadow masks and shadow-invariant color images as prior information to mitigate the dependence on one of the auxiliary tasks and improve the robustness of shadow removal. In addition, shadow synthesis can improve the performance of shadow removal by data augmentation [24], [49] or joint learning of shadow removal and generation [23], [31]–[34].

In this paper, we propose a novel decoupled multi-task shadow removal network that jointly learns shadow removal, shadow matte estimation, and shadow image reconstruction in one stage and one branch structure, with the help of the proposed constrained feature decoupling layer.

Comparison with Shadow Mask: Shadow masks are widely used for the shadow removal task to improve performance [21], [22], [24], [28], [31], [48]. The shadow mask is a binary image (0 or 1). Le and Samaras [22], [25] estimate shadow mattes (β) to remove shadows based on the shadow linear model ($I_{relit} = k \cdot I_s + b$) [13] and shadow image decomposition ($I_f = I_s \cdot (1-\beta) + I_{relit} \cdot \beta$) [47]. I_{relit} is the relit image and k, b are illumination parameters. The shadow matte β in [22], [25] is a grayscale image, which is a soft shadow mask (0-1). In this paper, our shadow matte (I_m) represents the value of the illumination compensation of shadow areas in an image. Our shadow matte (I_m) is the residual image between the shadow-free (I_f) and the shadow image (I_s), which is an RGB image (0-255). Compared to the shadow mask that only indicates the approximate shadow position and loses the shadow intensity information, our shadow matte provides clues to the shadow removal task in terms of shadow location and intensity, and avoids manually adjusting the ground truth of shadow mask [21]. In addition, obtaining the ground truth of our shadow matte requires only one step (by Eq. 1), while Le and Samaras [22], [25] obtain the shadow matte (β) requiring two steps.

III. Proposed Method

The illumination of a surface in non-shadow areas can be expressed as $L = L_a + L_d$, where L_a is the ambient illumination and L_d is the direct illumination [13]. The essence of shadow removal is to restore L_d in shadow areas. Given a shadow image I_s in source domain S and its corresponding shadow-free image I_f in target domain T, our goal is to learn the mapping $M : S \rightarrow T$, where $T = S + R$, and R is direct illumination compensation (also known as shadow matte I_m). We can formulate a shadow-free image as

$$I_f = I_s + I_m. \tag{1}$$

Modern learning-based methods usually directly learn the mapping M with a deep neural network (G) to restore a shadow-free image[3] $\tilde{I}_f$ that is close to the reference I_f, i.e, $\tilde{I}_f = G(I_s)$, which requires aggregating the features F learned by G into $\tilde{I}_f$ in the last layer of G. Notably, all channels of F are destined to represent I_f in the training of G. However, similar to the image decomposition principle of Eq. 1 in the image domain, we argue that a similar decomposition principle should also work in the feature space, where some channels of F represent I_m, and others represent I_s. We refer to this decomposition principle in the feature space as feature decomposition.

To this end, we embed feature decomposition in our shadow removal network. Specifically, we first propose a single-stage and decoupled multi-task network (Sec. III-A) that learns the features (F) decoupled in the channel dimension for three tasks, i.e., $\tilde{I}_f, \tilde{I}_m, \tilde{I}_s = G(I_s)$. Second, we illustrate how to capture the features F that are capable of representing I_m and I_s by feature transformation and refinement (Sec. III-B). Third, we design a feature decoupling layer (Sec. III-C) that can automatically select the feature channels of $\tilde{I}_m$ and $\tilde{I}_s$ from F, which are mutually exclusive. Finally, these decoupled features are aggregated into $\tilde{I}_f$, $\tilde{I}_m$, and $\tilde{I}_s$, supervised by multi-task joint loss functions (Sec. III-D).

A. Overall Network Architecture

Our DMTN (Fig. 3) takes a shadow image as input and simultaneously restores a shadow-free image, estimates a shadow matte, and reconstructs a shadow image. It consists of three steps. (1) **Feature extraction:** We apply a well pretrained VGG19 [50] for feature extraction and perform

[3]Throughout the paper, I denotes the original image from the dataset, and $\tilde{I}$ represents the image generated by a neural network.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3252271

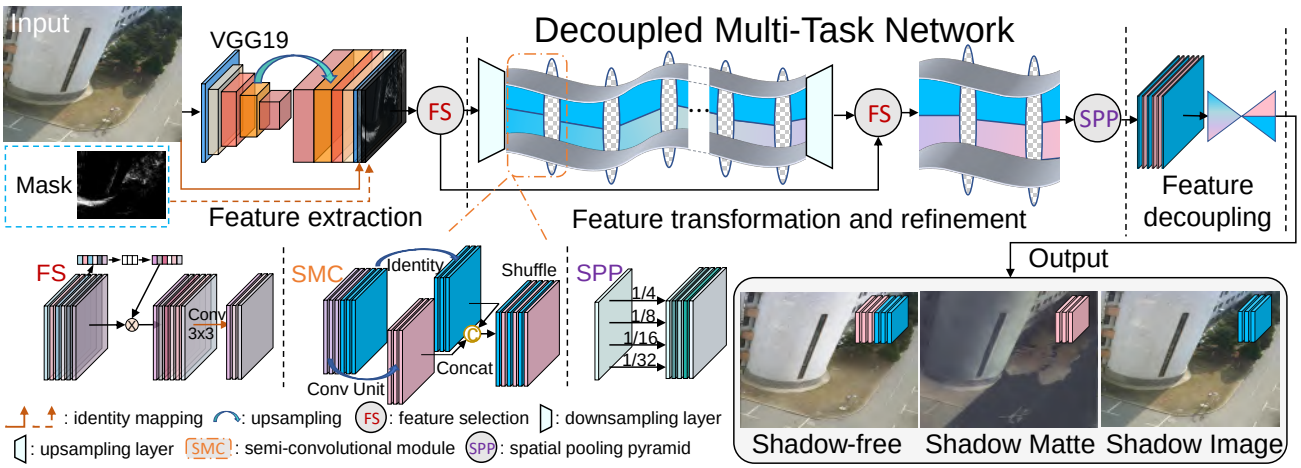

Fig. 3. Proposed decoupled multi-task network for shadow removal. (i) We select the optimal feature combination from the pretrained VGG19. (ii) We convert and refine some extracted features into shadow matte features, while retaining other features for shadow image reconstruction. (iii) We decouple the learned features into shadow-free, shadow matte, and reconstructed shadow images. Notably, the input shadow mask is not necessary, which an optional experimental setting (denoted as DMTN+Mask).

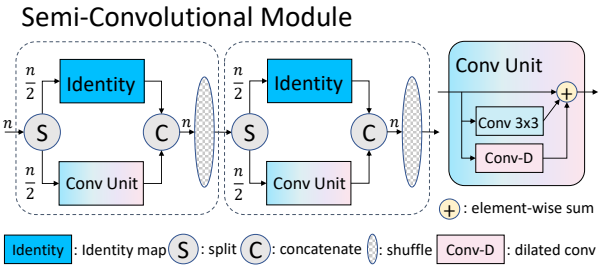

Fig. 4. Improved semi-convolution module for feature transformation and refinement. Half of the channels are processed to estimate the illumination compensation in the shadow areas; the other half are not processed, to preserve some features to represent shadow images. Multiple semi-convolution modules are cascaded to capture multiscale features for shadow removal.

3×3 convolution after the SE module [51] to automatically select global context and local detailed features [52], [53]. (2) **Feature transformation and refinement:** We adopt a coarse-to-fine structure to convert and refine some extracted features into shadow matte features, while retaining other features for shadow image reconstruction. (3) **Feature decoupling:** We decouple the learned features into three task domains by reassigning weights to feature channels.

B. Feature Transformation and Refinement

Because we extract features from I_s, some features should be allowed directly through the network to represent I_s, and some converted to represent I_m. Thus, we first introduce a semi-convolution (SMC) module that only processes half of the feature channels without processing the other half, and then concatenates and shuffles features (inspired by RealNVP [54] and ShuffleNet [55]), as shown in Fig. 4. However, the vanilla convolution module in ShuffleNet [55] has three cascaded convolutions (1×1, 3×3 depthwise [56], [57], and 1×1), which cannot accurately estimate shadow matte due to the absence of a large receptive field. Thus, we design a convolution unit including dilated and 3×3 convolution to model the direct illumination compensation to represent I_m. BatchNorm [58] in ShuffleNet [55] degrades shadow removal performance (see Table V(a)). Hence our DMTN uses adaptive normalization [59]. Then we adopt a coarse-to-fine structure consisting of multiple SMC modules (Fig. 3) to progressively convert the original shadow features to shadow matte features, while the remaining features are untouched to represent shadow images. Finally, we use a spatial pooling pyramid [60] to cope with shadows of various sizes.

The dilation rate of dilated convolution in the k-th ($k \geq 0$) SMC module is set to $2^{(k\%6)}$. At the coarse level, the feature map size is halved by a downsampling layer. At the refinement level, the features are refined to the original image resolution to improve the spatial accuracy of illumination compensation.

C. Feature Decoupling

After feature transformation and refinement using the introduced SMC module, we cannot identify which channels of the learned features F represent I_s, and which represent I_m. As shown in Fig. 5, we propose a constrained feature decoupling layer (CFDL) that can explicitly identify the feature channels belonging to each target task by weight decoupling to resolve the feature entanglement dilemma in multi-task learning.

1) Weight Constraint: First, we consider the most direct way to learn multiple tasks with one feature. In the last layer of the network, the learned features F (dim=[C,H,W]) are aggregated into RGB images (dim=[3,H,W]),

$$\begin{cases} \tilde{I}_s = \omega_s \times F, \\ \tilde{I}_m = \omega_m \times F, \\ \tilde{I}_f = \omega_f \times F, \end{cases} \tag{2}$$

where ω_s, ω_m, and ω_f are the parameters (dim=[3,C]) of 1×1 convolutions. Due to the absence of any constraint in Eq. 2, all channels of F are used for the three tasks.

Then, we try to introduce weight constraints to decouple the features in the channel dimension. Combining Eq. 1 and Eq. 2, we have $\omega_f \times F = \omega_s \times F + \omega_m \times F$, i.e.,

$$\omega_f = \omega_s + \omega_m. \tag{3}$$

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3252271

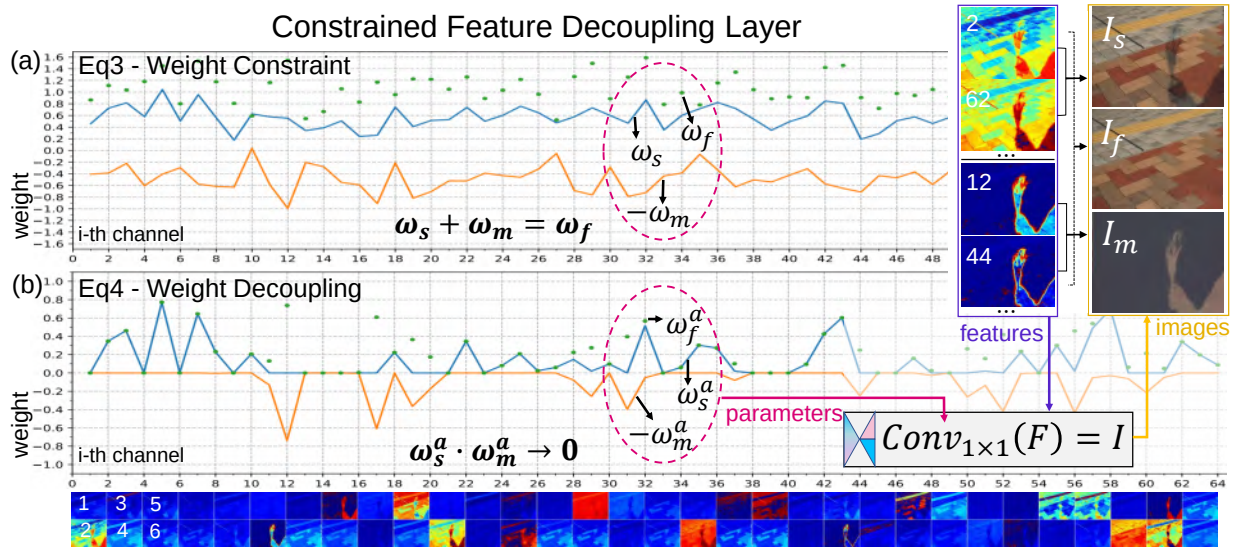

Fig. 5. Proposed constrained feature decoupling layer. (a) The features learned by our network are used for shadow image reconstruction and shadow matte estimation, and these two types of features are aggregated for shadow removal, but they are entangled; (b) By reassigning weights to channels using Eq. 4, these entangled features are decoupled and then aggregated into a shadow-free image, shadow matte, and reconstructed shadow image for multi-task collaborative learning. We use $-\omega_m$ instead of ω_m to clearly visualize all weights. Some magnified feature channel images show that the learned features have been decoupled. The figure shows the weights of the R channel only. Therefore, for some channels in Fig. 5(b) with weight values of 0, this does not mean that the channel is invalid because they are used for the aggregation of G or B channel (see Fig. 7).

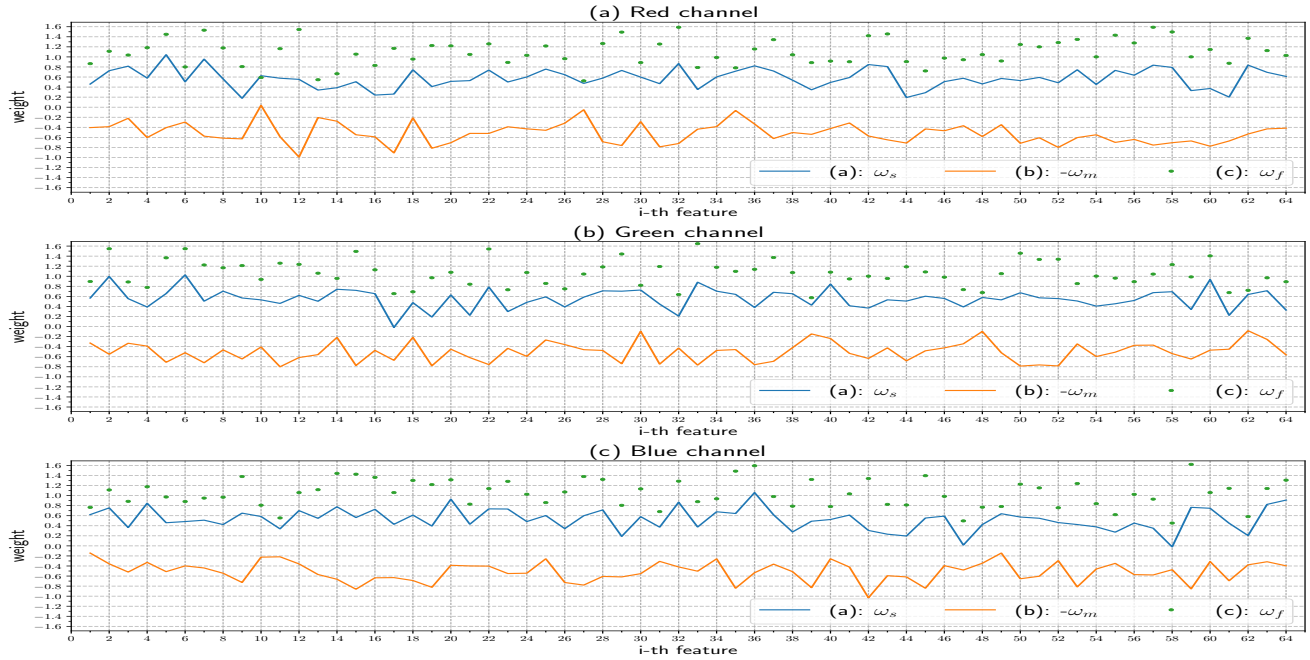

Fig. 6. Visualization of weight constraint (Eq. 3) for R, G, B channels. In (a)–(c), the visualization of the weights (ω_s, ω_m) does not clearly show which feature channels are used for shadow image reconstruction (ω_s), and which for shadow matte estimation (ω_m). In fact, almost every feature channel is used to learn the two tasks simultaneously. This means that feature entanglement still exists if only Eq. 3 is used to constrain these weights. We use $-\omega_m$ instead of ω_m to clearly visualize all weights.

Eq. 3 is the embodiment of the shadow image decomposition model (Eq. 1) in parameter space. However, the visualization of the Eq. 3 (see Fig. 6 or Fig. 5(a)) does not clearly show which feature channels are used for shadow image reconstruction, and which for shadow matte estimation. In fact, almost all channels are used to learn the two tasks simultaneously, which means that feature entanglement still exists if only Eq. 3 is used to constrain these weights. Thus, a new feature decoupling mechanism is needed.

2) Weight Decoupling: To decouple the features F learned by DMTN for multi-task learning, we expect to learn a parameter that can induce weight (ω_s, ω_m) decoupling by a weight adjustment function. Our motivation is to make the difference between ω_s and ω_m progressively larger in the channel dimension, and then to identify which target task the channel belongs to by comparing the magnitudes of the weights. Thus, we design a weight adjustment function (Eq. 4) with learnable parameters ω_b (dim=[3,C]) to induce weight

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3252271

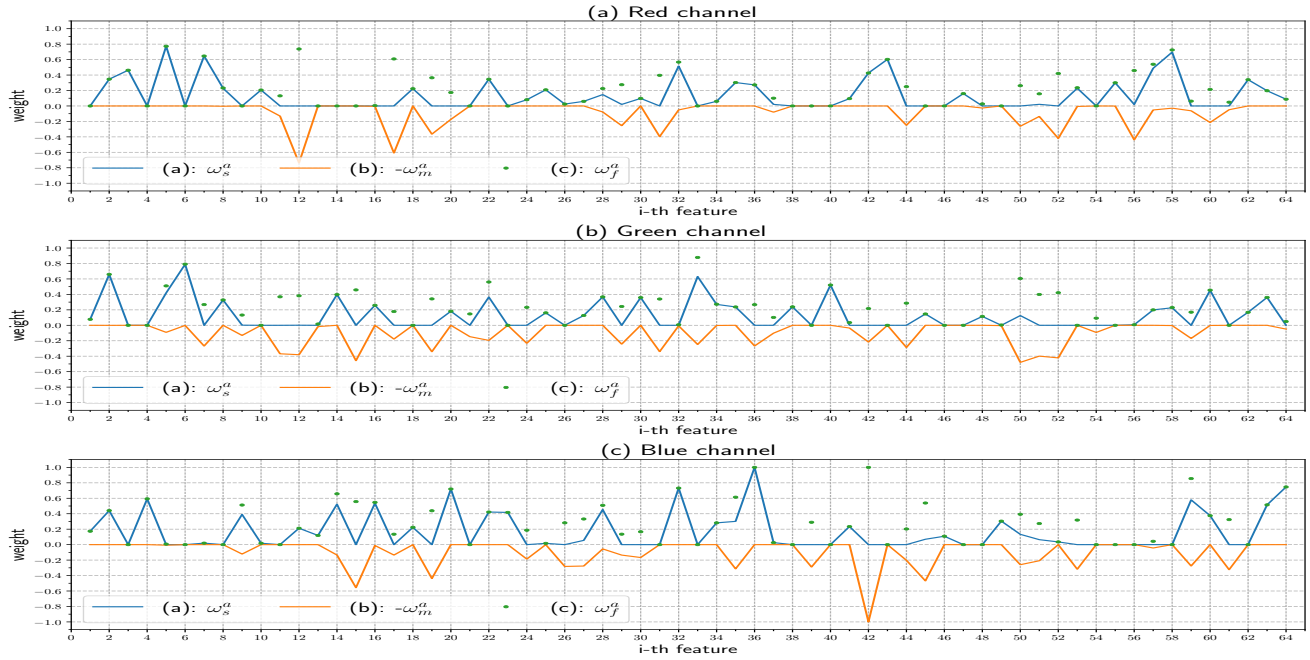

Fig. 7. Visualization of weight decoupling (Eq. 4) for R, G, B channels. In (a)–(c), almost each feature channel is used for only one task (shadow image reconstruction or shadow matte estimation) for each color channel (R, G, B), which shows that Eq. 4 can decouple features. These decoupled features are finally aggregated into a shadow-free image, shadow matte, and reconstructed shadow image. We use $-\omega_m^a$ instead of ω_m^a to clearly visualize all weights.

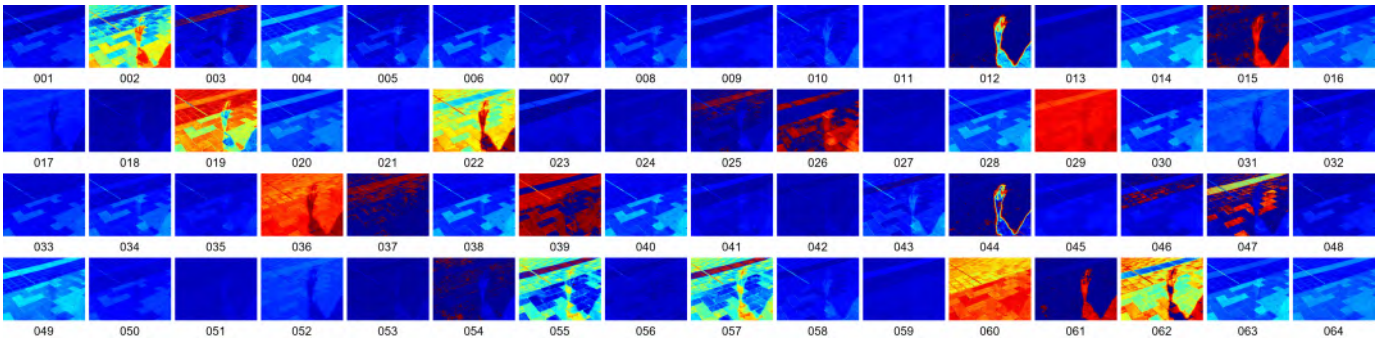

Fig. 8. Visualization of feature maps learned by DMTN with 64 hidden layers. By constrained feature decoupling layer (CFDL) and multi-task joint loss, some features can represent shadow images (I_s), and others can represent shadow matte (I_m). The weights corresponding to these feature channels are shown in Fig. 7(a)–(c).

decoupling,

$$\begin{cases} \omega_s^a = MedReLU_+(\omega_s \times (1+\omega_b)), \\ \omega_m^a = MedReLU_-(\omega_m \times (1-\omega_b)), \\ \omega_f^a = N(\omega_s^a) + N(\omega_m^a), \end{cases} \tag{4}$$

where ω_s^a, ω_m^a, and ω_f^a are the decoupled weights (dim=[3,C]), $MedReLU_{\pm}(x) = relu(x - median(x) \pm b)$, $N_{x \to z} : z = y/max(y), y = x - min(x)$. $median(x) \pm b$ is the adaptive median. In Eq. 4, if the weight of a channel is less than the median value in the channel dimension, the weight of the channel is set to zero by $MedReLU_-$, which means that the channel is not used to aggregate into $\tilde{I}_m$, and does not belong to the shadow matte estimation task. Considering that ω_s^a and ω_m^a may not be on the same scale after weight adjustment, we normalize the weights by N.

Fig. 7 (or Fig. 5(b)) shows the visualization of weight decoupling, i.e., $(\omega_s^a)_i \cdot (\omega_m^a)_i \to 0$, $i \in \{R, G, B\}$. The dimensions of $(\omega_s^a)_i$ are [1,C]. Notably, the weights in Fig. 7 come from a well-trained network. With the help of weight decoupling, we can achieve feature decoupling, i.e., some channels of F represent shadow images (I_m), and others represent shadow matte (I_m), as shown in the upper right corner of Fig. 5 (or Fig. 8).

We use the same learnable parameters b (dim=[3,1]) in $MedReLU_+$ and $MedReLU_-$ to balance the tasks of shadow image reconstruction and shadow matte estimation. b is initialized to zero. If $b = 0$, the values of half the channels are set to zero, and the numbers of channels belonging to $\tilde{I}_s$ and $\tilde{I}_m$ are equal. If $b > 0$, the number of channels of $\tilde{I}_s$ increases, and if $b < 0$, the number of channels of $\tilde{I}_m$ increases. Thus, b balances the number of channels used to learn the two tasks.

3) Theory Analysis for Weight Decoupling: We analyze why the weights are decoupled by Eq. 4. If only ℓ_1 loss is

considered, these learnable weights in Eq. 4 are updated as

$$\omega_{s_{k+1}} = \omega_{s_k} + \eta_s(1 + \omega_{b_k}), \tag{5}$$

$$\omega_{m_{k+1}} = \omega_{m_k} + \eta_m(1 - \omega_{b_k}), \tag{6}$$

$$\omega_{b_{k+1}} = \omega_{b_k} + \eta_b(\omega_{s_k} - \omega_{m_k}), \tag{7}$$

where η_s, η_m, and η_b are the step sizes. ω_s and ω_m are initialized using a normal distribution (N(0,1)), while ω_b are initialized to zero. If we subtract Eq. 6 from Eq. 5, we have

$$\Delta(\omega_s - \omega_m)_k = (\eta_s + \eta_m)\omega_{b_k} + (\eta_s - \eta_m), \tag{8}$$

where $\Delta(\omega_s - \omega_m)_k = \omega_{s_{k+1}} - \omega_{m_{k+1}} - (\omega_{s_k} - \omega_{m_k})$. If $\omega_s > \omega_m$, then $\omega_b > 0$ (Eq. 7), which leads to $\Delta(\omega_s - \omega_m) \uparrow$ (Eq. 8). In other words, ω_b makes the difference between ω_s and ω_m larger in the process of parameter updating. It is the same for $\omega_s < \omega_m$. Only those channels with weights greater than the median value are retained, while the others are discarded by $MedReLU_{\pm}$. Thus, ω_b can induce the weights to be decoupled, i.e., $\omega_s^a \cdot \omega_m^a \to 0$ in Fig. 5(b) (or Fig. 7).

4) Feature Aggregation: In Eq. 4, the weights $(\omega_s^a, \omega_m^a, \omega_f^a)$ are forced to be normalized to [0,1], which is actually a redundant limitation to aggregate the learned features F to images. This may limit the representation capability of the network, and thus we perform an adaptive linear transformation by learning the parameters α_i and β_i ($i \in \{s, m, f\}$, dim=[3,1,1]).

$$\begin{cases} \tilde{I}_s = \alpha_s \cdot \omega_s^a \times F + \beta_s, \\ \tilde{I}_m = \alpha_m \cdot \omega_m^a \times F + \beta_m, \\ \tilde{I}_f = \alpha_f \cdot \omega_f^a \times F + \beta_f. \end{cases} \tag{9}$$

α_i and β_i are initialized to 1 and 0, respectively.

D. Multi-task Collaborative Learning

The estimated shadow matte determines shadow areas and illumination compensation, while the reconstructed shadow image, as a constraint, preserves the features of the input image as much as possible. We adopt the ℓ_1 norm to constrain these two tasks,

$$\mathcal{L}_s = \mathbb{E}_{\tilde{I}_s, I_s}\left[\lambda_s \left\|\tilde{I}_s - I_s\right\|_1\right], \tag{10}$$

$$\mathcal{L}_m = \mathbb{E}_{\tilde{I}_m, I_m}\left[\lambda_m \left\|\tilde{I}_m - I_m\right\|_1\right]. \tag{11}$$

From Eq. 1, we obtain the ground truth of the shadow matte by $I_m = I_f - I_s$. For shadow removal, to restore a shadow-free image without artifacts and ensure the fidelity of non-shadow areas, we utilize the ℓ_1 norm to capture the low frequency, adversarial loss to capture high-frequency details [61], [62], and perceptual loss to improve perceptual quality [53], [63],

$$\mathcal{L}_f = \lambda_1 \mathcal{L}_{\ell_1} + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_{perc}. \tag{12}$$

Concatenating I_f and I_s into the discriminator D, the adversarial loss is

$$\mathcal{L}_{adv} = \mathbb{E}_{I_f, I_s}\left[\log\left[D(I_f, I_s)\right]\right] + \mathbb{E}_{\tilde{I}_f, I_s}\left[\log\left[1 - D(\tilde{I}_f, I_s)\right]\right]. \tag{13}$$

The perceptual loss is

$$\mathcal{L}_{perc} = \sum_{k=1}^{5} \mathbb{E}_{\tilde{I}_f, I_f}\left[\left\|\Phi_k(\tilde{I}_f) - \Phi_k(I_f)\right\|_1\right], \tag{14}$$

where Φ_k is the k-th activation map from the pretrained VGG19 [50]. Our final objective is

$$\arg\min_G \max_D (\mathcal{L}_s + \mathcal{L}_m + \mathcal{L}_f). \tag{15}$$

We empirically set $\lambda_s = 2550$, $\lambda_m = 255$, $\lambda_1 = 2550$, $\lambda_2 = 1$, $\lambda_3 = 1000$.

IV. Experiments

A. Implementation Details

DMTN was implemented with PyTorch and optimized with Adam (β_1=0.9 and β_2=0.999) [64]. In the feature transformation and refinement structure, the downsampling and upsampling layers were implemented using bilinear interpolation followed by 3×3 convolution. For training stability, we used historical images instead of the latest generated images to update the discriminator [65] with spectrum normalization [66], [67]. Identity parameter initialization [59] and random normal initialization were applied to initialize the generator and discriminator, respectively. The learning rates of the discriminator and generator were initially set to 1×10^{-4} and 5×10^{-5}, respectively. Similar to [24], we set the batch

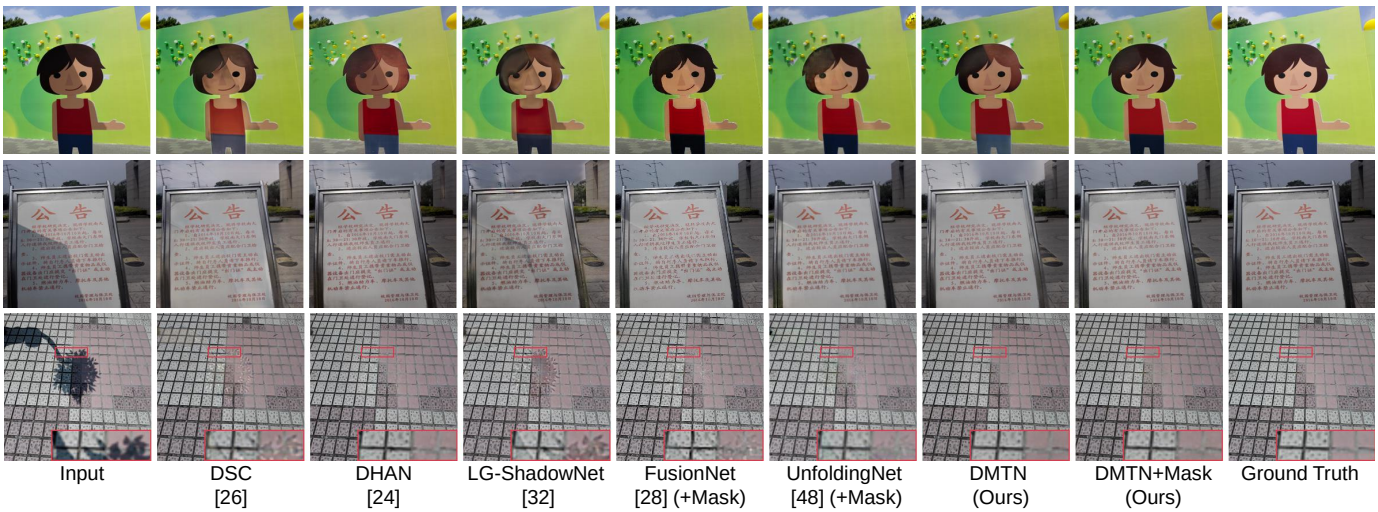

Fig. 9. Visual comparison results on ISTD dataset [21].

TABLE I

QUANTITATIVE COMPARISON RESULTS ON THE ISTD DATASET [21]. WE REPORT THE RMSE, SSIM AND PSNR IN THE SHADOW AREA, NON-SHADOW AREA, WHOLE IMAGE (ALL). "↑" INDICATES THAT LARGER VALUES ARE BETTER, WHILE "↓" INDICATES THAT LOWER VALUES ARE BETTER. BEST AND SECOND BEST RESULTS ARE **highlighted** AND UNDERLINED. THE RESULTS MARKED "∗", "¶" AND "‡" ARE REPORTED BY [21], [24] AND [30], RESPECTIVELY. DA REPRESENTS DATA AUGMENTATION BY DATA SYNTHESIS [24] FOR TRAINING. "-" REPRESENTS THAT THE TEST RESULTS ARE NOT AVAILABLE.

Method	RMSE(↓)			SSIM(↑)			PSNR(↑)		
	Shadow	Non-shadow	ALL	Shadow	Non-shadow	ALL	Shadow	Non-shadow	ALL
Yang [68] ∗	19.82	14.83	15.63	0.933	-	-	28.01	-	-
Guo [69] (TPAMI'12) ∗	18.95	7.46	9.30	0.964	0.966	0.920	27.76	26.44	23.08
Gong [16] (BMVC'14) ∗	14.98	7.29	8.53	0.973	0.972	0.926	30.14	26.98	24.71
ST-CGAN [21] (CVPR'18)	10.33	6.93	7.47	0.981	0.958	0.929	33.74	29.51	27.44
ARGAN [27] (ICCV'19) ‡	9.21	6.27	6.63	-	-	-	-	-	-
DSC [26] (TPAMI'19) ¶	9.48	6.14	6.67	0.967	-	-	33.45	-	-
RIS-GAN [29] (AAAI'20) ‡	9.21	6.27	6.63	-	-	-	-	-	-
DHAN [24] (AAAI'20)	8.14	6.04	6.37	0.983	-	-	34.50	-	-
DAD [70] (CVPR'20)	-	-	6.57	-	-	-	-	-	-
CANet [30] (ICCV'21)	8.86	6.07	6.15	-	-	-	-	-	-
LG-ShadowNet [32] (TIP'21)	10.23	5.38	6.18	0.979	0.967	0.936	31.53	29.47	26.62
DiNet [71] (TMM'22)	-	-	6.28	-	-	-	-	-	-
Ours (DMTN)	**7.36**	**5.05**	**5.43**	**0.989**	**0.973**	**0.958**	**35.29**	**31.25**	**29.04**
FusionNet [28] (CVPR'21+Mask)	7.77	5.56	5.92	0.975	0.880	0.945	34.71	28.61	27.19
UnfoldingNet [48] (AAAI'22+Mask)	7.87	4.72	5.22	0.987	0.978	0.960	**36.95**	31.54	29.85
BMNet [31] (CVPR'22+Mask)	7.60	4.59	5.02	0.988	0.976	0.959	35.61	32.80	30.28
Ours (DMTN+Mask)	**7.00**	**4.28**	**4.72**	**0.990**	**0.979**	**0.965**	35.83	**33.01**	**30.42**
DHAN [24] (AAAI'20+DA)	7.52	5.43	5.76	0.984	-	-	34.98	-	-
Ours (DMTN+DA)	**6.86**	**4.71**	**5.06**	**0.989**	**0.973**	**0.959**	**35.97**	**31.76**	**29.72**

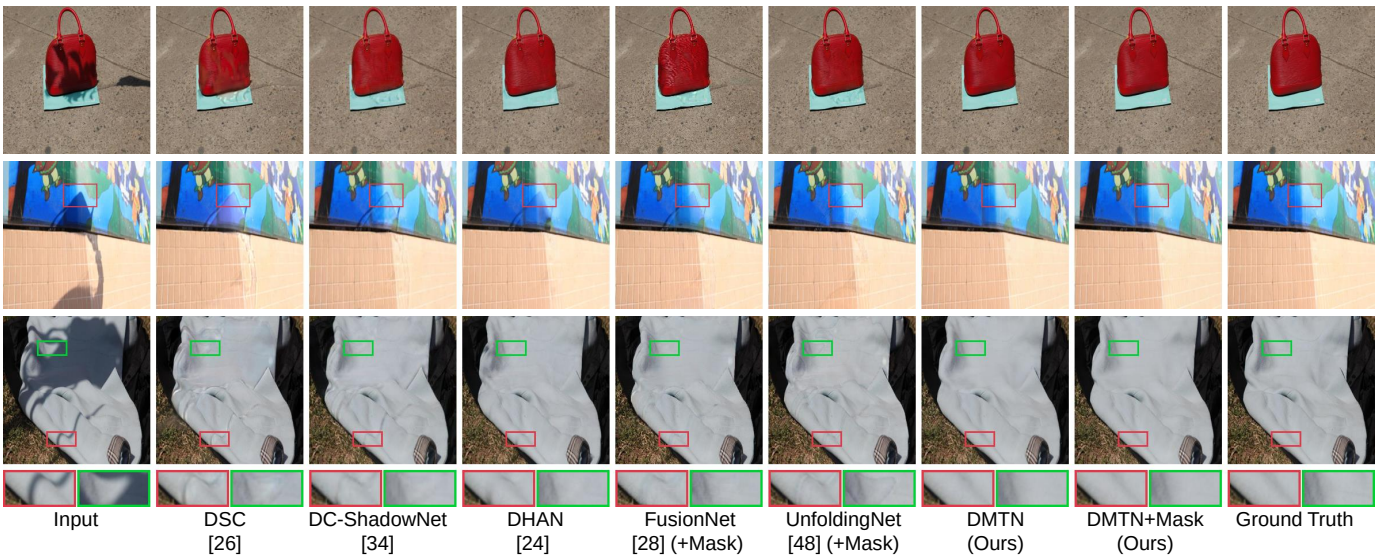

Fig. 10. Visual comparison results on SRD dataset [20].

size to 1 and randomly resize images for data augmentation. We trained DMTN in 90 epochs with data augmentation (total 150 epochs). Training and testing were performed on a single Nvidia GTX 3090.

B. Datasets and Evaluation Metrics

We evaluated our method on the widely used shadow removal datasets SRD [20], ISTD [21], and ISTD+ [22]. SRD [20] contains shadow images and shadow-free images (2,680 for training; 408 for testing). ISTD [21] contains shadow images, shadow masks, and shadow-free image triplets (1,330 for training; 540 for testing). ISTD+ [22] is an improved dataset to deal with the color inconsistencies in ISTD [21]. Cun *et al.* [24] trained a shadow matting GAN to synthesize the shadow dataset (ISTD+DA), which uses the shadow-free images in USR [23] and the shadow mask in ISTD [21]. The training set of ISTD+DA contains the original training set [21] and a synthetic training set [24]. We evaluated the SRD dataset using the shadow masks of Cun *et al.* [24]. The quantitative results of shadow removal were evaluated by root mean square error (RMSE), structural similarity (SSIM), and peak signal to noise ratio (PSNR).

C. Comparison with the State-of-the-art Methods

We compared DMTN with the following methods: (1) Yang [68], Guo [69], and Gong [16] are traditional methods; (2) DeShadowNet [20], DSC [26], and DHAN [24] are single-stage networks; (3) ST-CGAN [21], ARGAN [27], Param+M-Net [22], RIS-GAN [29], FusionNet [28], CANet [30], UnfoldingNet [48], and BMNet [31] are multi-stage networks; (4) Mask-ShadowGAN [23], DC-ShadowNet [34], and LG-ShadowNet [32] are unsupervised methods; (5) Param+M+D-

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3252271

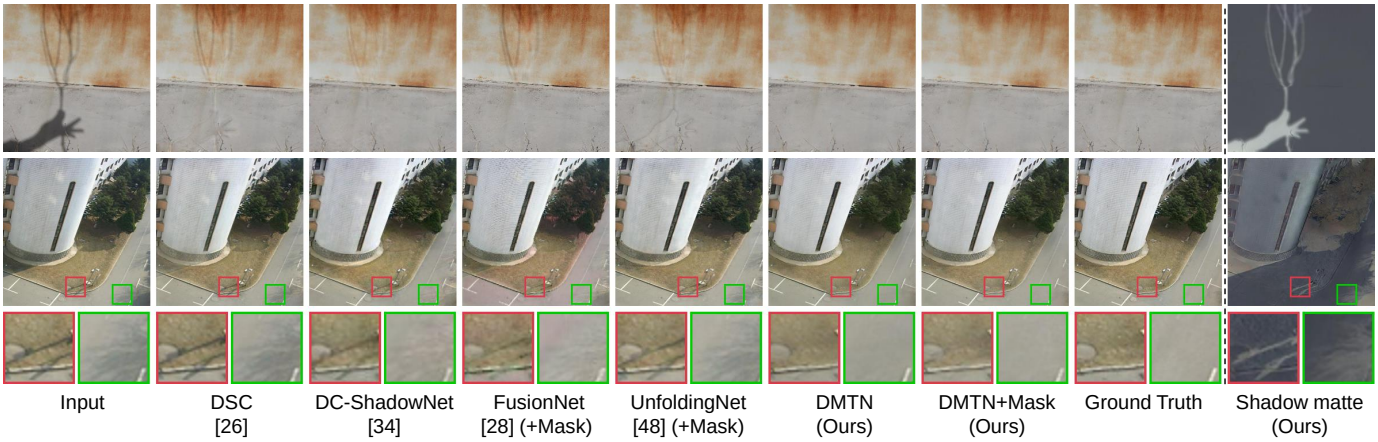

Fig. 11. Visual result of penumbra removal on SRD [20]. Last column is shadow matte estimated by our network.

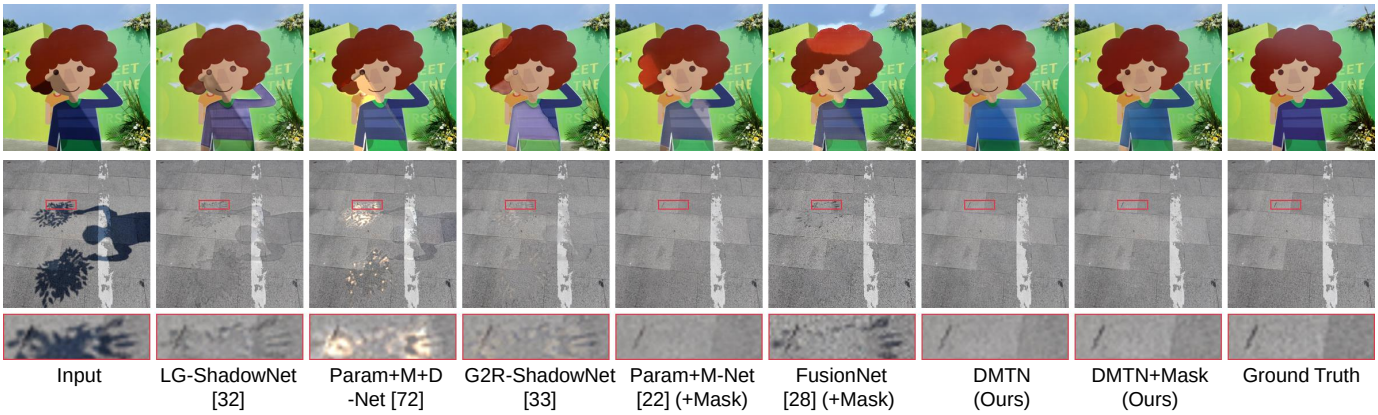

Fig. 12. Visual comparison results on ISTD+ dataset [22].

TABLE II
QUANTITATIVE COMPARISON RESULTS ON THE SRD DATASET [20] IN RMSE (THE LOWER THE BETTER). THE RESULTS MARKED ∗, ¶ AND ‡ ARE REPORTED BY [20], [24] AND [30], RESPECTIVELY. ALL REPRESENTS THE WHOLE IMAGE. BEST AND SECOND BEST RESULTS ARE **highlighted** AND UNDERLINED.

Method	Shadow	Non-shadow	ALL
Yang [68] (TIP'12) ∗	23.43	22.26	22.57
Guo [69] (TPAMI'12) ∗	29.89	6.47	12.60
Gong [16] (BMVC'14) ∗	19.58	4.92	8.73
DeShadowNet [20] (CVPR'17)	11.78	4.84	6.64
ARGAN [27] (ICCV'19) ‡	8.13	6.05	6.23
DSC [26] (TPAMI'19) ¶	10.89	4.99	6.23
RIS-GAN [29] (AAAI'20) ‡	8.09	6.02	6.17
DHAN [24] (AAAI'20)	8.94	4.80	5.67
DAD [70] (CVPR'20)	-	-	5.82
CANet [30] (ICCV'21)	7.82	5.88	5.98
DC-ShadowNet [34] (ICCV'21)	7.70	3.39	4.66
DiNet [71] (TMM'22)	-	-	5.38
Ours (DMTN)	**6.53**	**3.11**	**4.06**
FusionNet [28] (CVPR'21+Mask)	8.56	5.75	6.51
UnfoldingNet [48] (AAAI'22+Mask)	7.44	3.74	4.79
BMNet [31] (CVPR'22+Mask)	6.61	3.61	4.46
Ours (DMTN+Mask)	**5.92**	**3.03**	**3.82**

TABLE III
QUANTITATIVE COMPARISON RESULTS ON ISTD+ [22]. ALL REPRESENTS THE WHOLE IMAGE. "WEAKLY" REPRESENTS WEAKLY-SUPERVISED METHODS. BEST AND SECOND BEST RESULTS ARE **highlighted** AND UNDERLINED.

Method	Learning	Shadow	Non-shadow	ALL
Yang [68] (TIP'12)	traditional	24.7	14.4	16.0
Guo [69] (TPAMI'12)	traditional	22.0	3.1	6.1
Mask-ShadowGAN [23] (ICCV'19)	unsupervised	12.4	4.0	5.3
DC-ShadowNet [34] (ICCV'21)	unsupervised	10.3	3.5	4.6
LG-ShadowNet [32] (TIP'21)	unsupervised	9.7	3.4	4.4
Param+M+D-Net [72] (ECCV'20)	weakly	9.7	3.0	4.0
G2R-ShadowNet [33] (CVPR'21)	weakly	8.8	2.9	3.9
DeShadowNet [20] (CVPR'17)	supervised	15.9	6.0	7.6
ST-CGAN [21] (CVPR'18)	supervised	13.4	7.7	8.7
Param+M-Net [22] (ICCV'19+Mask)	supervised	7.9	3.1	3.9
FusionNet [28] (CVPR'21+Mask)	supervised	6.5	3.8	4.2
BMNet [31] (CVPR'22+Mask)	supervised	**5.6**	**2.5**	**3.0**
Ours (DMTN)	supervised	6.5	3.1	3.7
Ours (DMTN+Mask)	supervised	6.1	2.6	3.2

Net [72] and G2R-ShadowNet [33] are weakly-supervised methods; (6) DAD [70] and DiNet [71] are unified frameworks for superimposed image decomposition.

We present the quantitative comparison results with these state-of-the-art methods on the ISTD dataset [21] in Table I (visual results are shown in Fig. 9). For fair comparisons, the results of these methods are provided from the original papers as much as possible. If not specified, DMTN has 12 SMC modules in the coarse level, 2 SMC modules in the refinement level, and 64 hidden layers. First, we report the shadow removal performance of DMTN, which decreases the RMSE of shadow regions by 16.9% compared to CANet [30] (the lower the better). Second, similar to [22], [25], [28], [31], [48], [72] using shadow masks as auxiliary information to locate shadow areas, we concatenate the shadow masks with the extracted VGG features as the input to the feature selection module, denoted as DMTN+Mask. Using a shadow

mask to facilitate shadow removal reduces the RMSE in the whole image from 5.43 to 4.72. Finally, we explore training our network with synthetic shadow images for data augmentation, denoted as DMTN+DA. The results show that synthetic shadow images improve the performance of shadow removal, and DMTN+DA achieves the best RMSE (6.86) in the shadow regions. In addition to RMSE, we also report the quantitative comparison of PSNR and SSIM in Table I. These results show the superiority of our method.

We present the quantitative results on the SRD dataset [20] in Table II. Our method clearly achieves the best shadow removal performance. Fig. 10 shows that our network can effectively recover the illumination in the shadow regions by multi-task collaboration. We analyze why our network can achieve good visual results for shadow removal by visualizing the output of the shadow matte estimation task in the last column of Fig. 11. The shadow matte estimated by our network can accurately model direct illumination compensation, which is the key to removing penumbra and tiny shadows. These results validate the effectiveness of our multi-task learning.

Table III presents quantitative results on the ISTD+ dataset [22] (visual results are shown in Fig. 12). Our method achieves the second-best performance, with a slightly lower RMSE (0.2) than BMNet [31] in the whole image. BMNet [31] achieves the best performance by introducing shadow masks and shadow-invariant color images as prior information, which shows that more priors can improve the performance of single-image shadow removal.

D. Ablation Studies

We conducted ablation experiments to demonstrate the effectiveness of our method. Due to the simplified training settings, the quantitative results of ablation experiments differ from the comparison results described above.

TABLE IV
DMTN ARCHITECTURE ANALYSIS ON THE ISTD DATASET [21]. "DMTN/C=64/12-SMC/2-SMC" HAS 64 HIDDEN LAYERS, 12 SMC MODULES IN THE COARSE LEVEL AND 2 SMC MODULES IN THE REFINEMENT LEVEL. "W/O VGG" DENOTES THE VGG WITHOUT PRETRAINED WEIGHTS FOR FEATURE EXTRACTION. "W/O FS" REFERS TO THE 3×3 CONVOLUTION WITHOUT SE CHANNEL ATTENTION [51] FOR FEATURE SELECTION. "W/O SPP" DENOTES THE DMTN WITHOUT SPP MODULE IN FIG. 3.

Method/Channel/Coarse/Refine	Shadow	Non-Shadow	ALL
DMTN/C=64/4-SMC/2-SMC	7.53	5.71	6.00
DMTN/C=64/8-SMC/2-SMC	7.99	5.60	5.98
DMTN/C=64/16-SMC/2-SMC	8.22	5.73	6.12
DMTN/C=64/12-SMC/1-SMC	7.56	5.49	5.82
DMTN/C=64/12-SMC/3-SMC	7.82	5.72	6.05
DMTN/C=64/12-SMC/4-SMC	7.30	5.56	5.84
DMTN/C=64/14-SMC/0-SMC	7.62	5.61	5.93
DMTN/C=64/0-SMC/14-SMC	8.01	5.90	6.23
DMTN/C=16/12-SMC/2-SMC	9.02	6.41	6.82
DMTN/C=32/12-SMC/2-SMC	8.45	5.78	6.20
DMTN/C=64/12-SMC/2-SMC (**Ours**)	7.22	5.50	5.77
DMTN/C=128/12-SMC/2-SMC	**7.07**	**5.30**	**5.58**
DMTN/C=256/12-SMC/2-SMC	7.20	5.33	5.62
w/o VGG/C=64/12-SMC/2-SMC	8.24	6.22	6.54
w/o FS/C=64/12-SMC/2-SMC	7.58	5.64	5.94
w/o SPP/C=64/12-SMC/2-SMC	7.43	5.59	5.88

TABLE V
QUANTITATIVE COMPARISON WITH SHUFFLENET [55], [73] ON THE ISTD DATASET [21] IN RMSE. SHUFFLENETV1 [73] AND SHUFFLENETV2 [55] USE BATCHNORM [58] AS THE DEFAULT SETTING, WHILE OUR DMTN USES ADAPTIVE NORMALIZATION ("ADANORM") [59]. (A) QUANTITATIVE COMPARISON WITH SHUFFLENET. "G=4" REPRESENTS THAT THE NUMBER OF GROUPS IS 4 IN GROUP CONVOLUTION. SHUFFLENETV1, SHUFFLENETV2, AND DMTN HAVE 64 HIDDEN LAYERS. (B) QUANTITATIVE COMPARISON OF RESOURCE EFFICIENCY AND PERFORMANCE. "C=44" DENOTES 44 HIDDEN LAYERS. "MACS" DENOTES MULTIPLY–ACCUMULATE OPERATION. RESOURCE EFFICIENCY (PARAMS AND MACS) IS ANALYZED BY THOP[4].

(a) Method	Shadow	Non-Shadow	ALL
ShuffleNetV1 [73]/g=1	12.36	7.54	8.30
ShuffleNetV1 [73]/g=2	13.19	8.14	8.93
ShuffleNetV1 [73]/g=4	12.21	7.18	7.97
ShuffleNetV1 [73]/g=8	11.85	8.54	9.06
ShuffleNetV1 [73]/g=1/AdaNorm	8.00	5.94	6.27
ShuffleNetV1 [73]/g=2/AdaNorm	8.28	5.91	6.28
ShuffleNetV1 [73]/g=4/AdaNorm	8.70	5.97	6.40
ShuffleNetV1 [73]/g=8/AdaNorm	8.10	6.06	6.38
ShuffleNetV2 [55]	12.32	7.44	8.21
ShuffleNetV2 [55]/AdaNorm	8.27	6.03	6.38
Ours (DMTN)	**7.22**	5.50	**5.77**

(b) Method	Shadow	Non-Shadow	ALL	Params (M)	MACs (G)
ShuffleNetV2 [55]/c=44	9.13	6.16	6.63	21.44	58.42
Ours (DMTN/c=32)	8.45	5.78	6.20	21.35	48.81
ShuffleNetV2 [55]/c=80	8.08	5.59	5.98	22.56	104.06
Ours (DMTN/c=64)	7.22	5.50	5.77	22.83	93.65
ShuffleNetV2 [55]/c=184	7.59	5.58	5.90	27.89	306.12
Ours (DMTN/c=128)	**7.07**	5.30	**5.58**	27.90	230.22

The results in Table IV show that the coarse-to-fine structure is effective and essential for shadow removal, probably because the encoder-decoder structure can capture the global context to locate shadows, and the image processing network [24], [41], [59] at the original image resolution is conducive to restoring spatially accurate details. In addition, we discuss our network scalability for shadow removal by changing the number of hidden layers. The results show that blindly increasing parameters cannot further improve shadow removal performance. In addition, we conducted ablation analysis on feature extraction and selection (w/o VGG and w/o FS in Table IV). Notably, we used the original SE module [51] for feature selection. It is feasible to explore more complex attention modules for performance improvement, e.g., Expansion-Squeeze-Excitation attention [74]–[76], effective SE module [77], and Multi-Axis attention [42].

Our SMC module was improved based on the Shuffle module [55], [73]. There are two differences between SMC and Shuffle module [55], [73]: 1) the Shuffle module [55], [73] uses BatchNorm [58], while our SMC module uses adaptive normalization [59]; 2) the convolution unit of Shuffle module uses three cascaded convolutions (1×1, 3×3 depth-wise [56], [57], and 1×1), while our SMC module adopts two parallel convolutions (dilated and 3×3). For a fair comparison, both ShuffleNet and our DMTN have 12 modules in the coarse level and 2 modules in the refinement level. Table V(a) shows that adaptive normalization [59] can improve the performance of shadow removal. Table V(b) provides

[4]https://github.com/Lyken17/pytorch-OpCounter.

TABLE VI

ANALYSIS OF THE RELATIONSHIP BETWEEN THE QUALITY OF SHADOW MATTES AND SHADOW REMOVAL PERFORMANCE ON THE ISTD DATASET [21]. IN THIS ABLATION EXPERIMENT, WE REMOVED THE ADVERSARIAL AND PERCEPTUAL LOSSES (I.E., W/O ($\mathcal{L}_{adv}, \mathcal{L}_{perc}$)), AND USED ONLY THE ℓ_1 NORM. BY BLURRING THE GROUND TRUTH OF SHADOW MATTES, WE CAN CONTROL THE QUALITY OF THE PREDICTED SHADOW MATTES. WE CALCULATE "RMSE-SHADOW MATTE" USING THE PREDICTED SHADOW MATTE AND THE UNBLURRED GROUND TRUTH OF SHADOW MATTE.

Loss (w/o ($\mathcal{L}_{adv}, \mathcal{L}_{perc}$))	Shadow matte	RMSE-Shadow	RMSE-Non-Shadow	RMSE-ALL	RMSE-Shadow matte
Only shadow removal	0×0 mean blur	8.35	6.31	6.63	-
Multi-task	0×0 mean blur	8.24	5.91	6.28	20.11
Multi-task	3×3 mean blur	8.72	5.73	6.20	20.44
Multi-task	5×5 mean blur	8.09	5.85	6.20	20.56
Multi-task	7×7 mean blur	8.17	5.88	6.24	20.81
Multi-task	11×11 mean blur	8.35	6.02	6.39	21.13
Multi-task	13×13 mean blur	8.26	6.05	6.40	21.22

TABLE VII

SMC MODULE ANALYSIS ON THE ISTD DATASET [21]. (A) CHANNEL PARTITION ABLATION ANALYSIS IN THE SMC MODULE (B) CONVOLUTION UNIT SELECTION IN THE SMC MODULE. AS SHOWN IN FIG. 4, "CONV UNIT" DENOTES PARALLEL DILATED AND 3×3 CONVOLUTION, "CONV 3×3" DENOTES 3×3 CONVOLUTION, "IDENTITY" DENOTES IDENTITY MAP, "CONV-D" DENOTES DILATED CONVOLUTION, AND "/" DENOTES SPLITTING AND PROCESSING FEATURE CHANNELS. "MACS" IS MULTIPLY–ACCUMULATE OPERATION. RESOURCE EFFICIENCY (PARAMS AND MACS) IS ANALYZED BY THOP.

(a) SMC	Shadow	Non-Shadow	ALL	Δ	Params (M)	MACs (G)
Conv Unit(5/5)	7.46	5.47	5.79	-	25.60	132.60
Conv Unit(4/5)	7.46	5.50	5.81	-0.02	24.25	113.65
Conv Unit(3/4)	7.34	**5.43**	**5.73**	**+0.08**	23.98	109.89
Conv Unit(2/3)	7.73	5.62	5.95	-0.22	23.53	103.50
Conv Unit(3/5)	7.82	5.45	5.83	**+0.12**	23.21	98.98
Conv Unit(1/2)	**7.22**	5.50	5.77	**+0.06**	22.83	93.65
Conv Unit(2/5)	7.71	5.48	5.83	-0.06	22.49	88.86
Conv Unit(1/3)	7.67	5.53	5.86	-0.03	22.30	86.24
Conv Unit(1/4)	7.61	5.79	6.07	-0.21	22.14	83.89
Conv Unit(1/5)	8.73	5.70	6.18	-0.11	**22.05**	**82.59**

(b) SMC	Shadow	Non-Shadow	ALL	Params (M)	MACs (G)
Identity/Conv 3×3	8.03	5.92	6.25	**22.37**	**87.13**
Identity/Conv-D	8.02	5.58	5.96	**22.37**	**87.13**
Identity/Conv Unit (**Ours**)	**7.22**	5.50	**5.77**	22.83	93.65
Conv 3×3/Conv Unit	7.62	5.55	5.87	23.29	100.18
Conv 1×1/Conv Unit	7.60	5.54	5.86	22.88	94.42
ALL channels	7.46	5.47	5.79	25.60	132.60

TABLE VIII

CONSTRAINED FEATURE DECOUPLING LAYER (SEE EQ. 4) ANALYSIS ON THE ISTD DATASET [21]. "W/O LINEAR" DENOTES THE FEATURE DECOUPLING WITHOUT LINEAR TRANSFORMATION IN EQ. 9.

CFDL	Shadow	Non-Shadow	ALL
Eq. 2	8.02	5.50	5.90
Eq. 2, Eq. 3	7.45	5.63	5.91
w/o b	7.46	5.40	5.73
w/o median	7.42	5.50	5.80
w/o adaptive median	8.11	**5.27**	**5.72**
ReLU→LeakyReLU	7.55	5.49	5.82
w/o normalize	7.74	5.59	5.93
w/o linear	7.62	5.52	5.85
Eq. 4, Eq. 9 (Ours)	**7.22**	5.50	5.77

TABLE IX

MULTI-TASK JOINT LOSS FUNCTIONS ANALYSIS ON THE ISTD DATASET [21]. "1×1 CONV" DENOTES AGGREGATING n DIMENSIONAL FEATURES INTO RGB IMAGES BY 1×1 CONVOLUTION.

Loss	CFDL	Shadow	Non-Shadow	ALL
w/o ($\mathcal{L}_s, \mathcal{L}_m$)	1×1 conv	7.74	5.68	6.00
w/o $\mathcal{L}_{\ell_1}$	Eq. 4, Eq. 9	7.56	5.62	5.93
w/o $\mathcal{L}_{adv}$	Eq. 4, Eq. 9	8.10	5.52	5.93
w/o $\mathcal{L}_{perc}$	Eq. 4, Eq. 9	8.67	6.02	6.44
$\mathcal{L}_s, \mathcal{L}_m, \mathcal{L}_f$ (Ours)	Eq. 4, Eq. 9	**7.22**	**5.50**	**5.77**

comparisons with ShuffleNet [55], [73] in terms of network complexity and accuracy, which demonstrates the effectiveness of our improved SMC module. In addition, we conducted channel partition ablation experiments on the SMC module. The results in Table VII(a) show that processing half of the feature channels performs better, which is probably due to sufficient feature mixing. We observed two interesting points: 1) not all scores drop, and some even improve when the number of convolved channels decreases (5/5 1/2); 2) the scores are getting worse (1/2 1/5). This shows that it is beneficial to pass some channels, but convolving a few channels leads to significant performance degradation (e.g., Conv Unit 1/5). In Table VII(b), we selected a setting of convolution unit that can balance performance and parameters. Compared with full channel convolution ("ALL channels" in Table VII(b)), our SMC module has fewer parameters and computational costs without performance degradation.

We analyzed the relationship between the quality of shadow mattes and shadow removal performance in Table VI. By blurring the ground truth of shadow mattes, we can control the quality of the predicted shadow mattes. We found that supervised training with slightly blurred shadow mattes (e.g., 5×5 mean blur) results in better RMSE (6.28 $\rightarrow$ 6.20), while overly inaccurate shadow mattes degrade the shadow removal performance (e.g., 13×13 mean blur). This may be due to slight changes of ambient light during image capture, where illumination noise is unavoidable (Le and Samaras [22] solved this issue by correcting color inconsistency in ISTD [21]). Therefore, there is a difference between shadow and shadow-free images, and slightly blurred shadow mattes can suppress noise interference.

In Table VIII, the RMSE in the shadow regions, with values from 8.02 (Eq. 2) to 7.22, shows that feature decoupling can facilitate shadow matte estimation and better restore illumination in shadow areas. Our CFDL reduces the RMSE in shadow regions from 8.11 (w/o adaptive median) to 7.22, which shows that the adaptive median can better balance the tasks of shadow matte estimation and shadow image reconstruction. The ablation results in Table IX verify the effectiveness of our multi-task joint loss functions.

TABLE X
AVERAGE INFERENCE TIME ON A NVIDIA GTX 3090 WITH THE RESOLUTION OF 512×512.

Method	Param+M-Net [22] (+Mask)	DHAN [24]	FusionNet [28] (+Mask)	CANet [30]	Ours (DMTN)	Ours (DMTN+Mask)
Inference Time	**0.0521**	0.1039	0.0902	0.1151	0.0844	0.0846

TABLE XI
RESOURCE EFFICIENCY AND PERFORMANCE ANALYSIS BY THOP. ALL REPRESENTS THE WHOLE IMAGE. "MACS"[5] DENOTES MULTIPLY–ACCUMULATE OPERATION.

Multi-branch (ISTD)	Shadow	Non-Shadow	ALL	Params (M)	MACs (G)
DHAN [24]	8.14	6.04	6.37	21.75	131.47
FusionNet [28] (+Mask)	7.77	5.56	5.92	141.84	83.05
Ours (DMTN-32)	8.31	5.38	5.86	**21.35**	**63.73**
Ours (DMTN-64)	7.36	5.05	5.43	22.83	122.30
Ours (DMTN-64+Mask)	**7.00**	**4.28**	**4.72**	22.83	122.34

Multi-stage (SRD)	Shadow	Non-Shadow	ALL	Params (M)	MACs (G)
CANet [30]	7.82	5.88	5.98	236.12	247.38
Ours (DMTN)	**6.53**	**3.11**	**4.06**	**22.83**	**122.30**
FusionNet [28] (+Mask)	8.56	5.75	6.51	141.84	**83.05**
Ours (DMTN+Mask)	**5.92**	**3.03**	**3.82**	**22.83**	122.34

Multi-stage (ISTD+)	Shadow	Non-Shadow	ALL	Params (M)	MACs (G)
Param+M-Net [22] (+Mask)	7.9	3.1	3.9	141.18	**39.87**
Ours (DMTN)	6.5	3.1	3.7	**22.83**	122.30
Ours (DMTN+Mask)	**6.1**	**2.6**	**3.2**	**22.83**	122.34

Table XI shows the comparison results of resource efficiency and performance. DMTN with 32 hidden layers outperforms the current state-of-the-art single-stage shadow removal network, DHAN [24], with double 64 hidden layers (dual branch) in the whole image. Compared with the multi-stage methods [22], [28], [30], [72], DMTN achieves higher performance with fewer parameters.

V. DISCUSSION AND LIMITATIONS

A. Limitations

Our method sometimes fails to completely remove shadows, as shown in Fig. 13. In the first row (on ISTD [21]), pixel values of non-shadow areas are changed due to inaccurate localization of shadow areas, which can be improved by using shadow masks as additional inputs. The second row (on SRD [20]) shows a case where the color in the shadow areas is incorrectly restored. We speculate that this case may be due to the long-tailed distribution of the dataset.

Table X shows that the inference time of our method is competitive with the SOTA methods. The run speed of our method is worse than that of Param+M-Net [22], which may be due to the fact that the features in our network are convolved at a higher resolution (Param+M-Net [22] uses 8 downsampling layers, while our network uses only one downsampling layer at the coarse level).

[5]When MACs evaluated by THOP, the default resolution is 224×224, but this resulted in Param+M-Net [22] and FusionNet [28] not working. Therefore, for all methods in Table XI, the MACs are calculated using images with the resolution of 256×256.

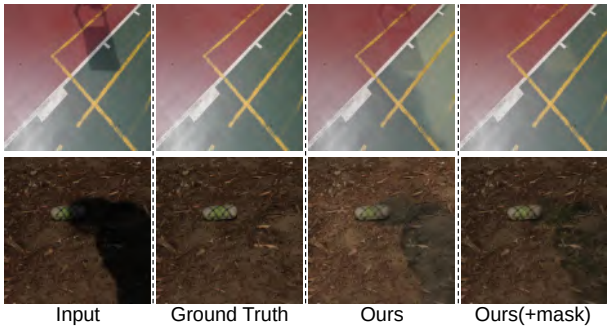

Fig. 13. Failure cases of our method.

B. Shadow Mask and Shadow Matte

We analyzed the relationship, advantages and disadvantages between the shadow mask and our matte. The input shadow mask facilitates the estimation of the shadow matte (Fig. 13), while our shadow matte is grayed and binarized to obtain the shadow mask [21]. Shadow matte is better for the recovery of shadow areas, while shadow mask is better for the fidelity of non-shadow areas. Compared to the shadow mask, our shadow matte can handle tiny shadows and penumbras, and avoid manually adjusting the ground truth of shadow mask [21]. But, our shadow matte causes noise interference in non-shadow areas, which can be observed in our ablation experiments (Table VI). Thus, shadow matte and shadow mask are a trade-off for shadow removal.

Compared to the widely used shadow mask, both our shadow matte and other mattes [20], [22], [25] can cope with penumbras. However, our shadow matte can avoid removing shadow in the log domain [14], [20] and estimating illumination parameters for shadow removal by a multi-stage network [22], [25]. In addition, our shadow matte is consistent with the residual image, facilitating the integration of shadow removal into the unified residual learning framework widely used for other image restoration tasks [29], [34], [36]–[42].

C. Self-shadow Removal

Although the training set of the SRD dataset [20] does not contain self-shadows and occluded objects, our results in Fig. 14 show a possibility to remove all shadows (umbra, penumbra, and self-shadow). Constructing a cast shadow removal dataset [20], [21] is simple, and shadow synthesis techniques [24], [49] are readily available, which provides technical support for generating a very large shadow removal dataset. Our results illustrate the feasibility that a network trained on a dataset containing only cast-shadow images can be used for self-shadow removal. In fact, due to the difficulty of constructing paired self-shadow removal datasets, our exploration points to a possible path for single-image shadow removal to be grounded in practice.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3252271

Fig. 14. Visual results of self-shadow removal. We collected a self-shadow removal dataset (SSRD) that contains 86 images without the ground truth of shadow-free images due to the presence of self-shadows. DHAN [24] and DMTN are pretrained on SRD dataset [20].

VI. Conclusion

In this paper, we proposed a decoupled multi-task shadow removal network to jointly learn three tasks in a single-stage pipeline: shadow removal, shadow matte estimation, and shadow image reconstruction. Our core design was an improved semi-convolution module for feature transformation and a constrained feature decoupling layer that decoupled the learned features into these three task domains by reassigning weights to feature channels, which is completely different from current methods using multi-stage or multi-branch structures for multi-task shadow removal. A theoretical analysis demonstrated the effectiveness of our decoupling mechanism, and comprehensive experiments illustrated the effectiveness and superiority of our method. In fact, the shadow matte estimated by our network was the residual image between the target and the input image; estimating the residual image has been widely used in the field of image restoration. Therefore, our method can potentially be applied to other image restoration tasks, e.g., highlight removal, low-light image enhancement, desnowing, and deraining.

References

[1] J. Stander, R. Mech, and J. Ostermann, "Detection of moving cast shadows for object segmentation," *IEEE TMM*, vol. 1, no. 1, pp. 65–76, 1999.

[2] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE TPAMI*, vol. 25, no. 10, pp. 1337–1342, 2003.

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

[4] W. Zhang, X. Z. Fang, X. K. Yang, and Q. J. Wu, "Moving cast shadows detection using ratio edge," *IEEE TMM*, vol. 9, no. 6, pp. 1202–1214, 2007.

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

[6] A. Sanin, C. Sanderson, and B. C. Lovell, "Improved shadow removal for robust person tracking in surveillance scenarios," in *Proc. ICPR*, 2010, pp. 141–144.

[7] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri, "Automatic Shadow Detection and Removal from a Single Image," *IEEE TPAMI*, vol. 38, no. 3, pp. 431–446, 2016.

[8] W. Zhang, X. Zhao, J. M. Morvan, and L. Chen, "Improving Shadow Suppression for Illumination Robust Face Recognition," *IEEE TPAMI*, vol. 41, no. 3, pp. 611–624, 2019.

[9] J. Shin, H. Park, and J. Paik, "Region-based dehazing via dual-supervised triple-convolutional network," *IEEE TMM*, vol. 24, pp. 245–260, 2021.

[10] Y. Zhong, X. Liu, D. Zhai, J. Jiang, and X. Ji, "Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon," in *Proc. CVPR*, 2022, pp. 15 345–15 354.

[11] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew, "On the removal of shadows from images," *Proc. CVPR*, vol. 28, no. 1, pp. 59–68, 2005.

[12] T. P. Wu, C. K. Tang, M. S. Brown, and H. Y. Shum, "Natural shadow matting," *ACM TOG*, vol. 26, no. 2, pp. 8—-es, 2007.

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

[14] E. Arbel and H. Hel-Or, "Shadow removal using intensity surfaces and texture anchor points," *IEEE TPAMI*, vol. 33, no. 6, pp. 1202–1216, 2011.

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

[16] H. Gong and D. Cosker, "Interactive shadow removal and ground truth for variable scene categories," in *Proc. BMVC*, M. F. Valstar, A. P. French, and T. P. Pridmore, Eds., 2014.

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

[18] G. D. Finlayson, M. S. Drew, and C. Lu, "Entropy minimization for shadow removal," *IJCV*, vol. 85, no. 1, pp. 35–57, 2009.

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

[20] L. Qu, J. Tian, S. He, Y. Tang, and R. W. Lau, "DeshadowNet: A multi-context embedding deep network for shadow removal," in *Proc. CVPR*, 2017, pp. 2308–2316.

[21] J. Wang, X. Li, and J. Yang, "Stacked Conditional Generative Adversarial Networks for Jointly Learning Shadow Detection and Shadow Removal," in *Proc. CVPR*, 2018, pp. 1788–1797.

[22] H. Le and D. Samaras, "Shadow removal via shadow image decomposition," in *Proc. ICCV*, 2019, pp. 8577–8586.

[23] X. Hu, Y. Jiang, C. W. Fu, and P. A. Heng, "Mask-shadowGAN: Learning to remove shadows from unpaired data," in *Proc. ICCV*, 2019, pp. 2472–2481.

[24] 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. 10 680–10 687.

[25] H. Le and D. Samaras, "Physics-based Shadow Image Decomposition for Shadow Removal," *IEEE TPAMI*, no. 01, p. 1, 2021.

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

[27] B. DIng, C. Long, L. Zhang, and C. Xiao, "ARGAN: Attentive recurrent generative adversarial network for shadow detection and removal," in *Proc. ICCV*, 2019, pp. 10 212–10 221.

[28] L. Fu, C. Zhou, Q. Guo, F. Juefei-Xu, H. Yu, W. Feng, Y. Liu, and S. Wang, "Auto-Exposure Fusion for Single-Image Shadow Removal," in *Proc. CVPR*, 2021, pp. 10 566–10 575.

[29] 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*, vol. 34, no. 07, 2020, pp. 12 829–12 836.

[30] Z. Chen, C. Long, L. Zhang, and C. Xiao, "CANet: A Context-Aware Network for Shadow Removal," in *Proc. ICCV*, 2021, pp. 4723–4732.

[31] Y. Zhu, J. Huang, X. Fu, F. Zhao, Q. Sun, and Z.-J. Zha, "Bijective mapping network for shadow removal," in *Proc. CVPR*, 2022, pp. 5627–5636.

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

[33] Z. Liu, H. Yin, X. Wu, Z. Wu, Y. Mi, and S. Wang, "From Shadow Generation to Shadow Removal," in *Proc. CVPR*, 2021, pp. 4925–4934.

[34] 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. ICCV*, 2021, pp. 5007–5016.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[36] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE TIP*, vol. 26, no. 7, pp. 3142–3155, 2017.

[37] H. Zhang, Y. Dai, H. Li, and P. Koniusz, "Deep stacked hierarchical multi-patch network for image deblurring," in *Proc. CVPR*, 2019.

[38] K. Jiang, Z. Wang, P. Yi, C. Chen, B. Huang, Y. Luo, J. Ma, and J. Jiang, "Multi-scale progressive fusion network for single image deraining," in *Proc. CVPR*, 2020, pp. 8346–8355.

[39] S. Anwar and N. Barnes, "Densely residual laplacian super-resolution," *IEEE TPMI*, vol. 44, no. 3, pp. 1192–1204, 2020.

[40] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE TPMI*, vol. 43, no. 7, pp. 2480–2495, 2020.

[41] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M. H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proc. CVPR*, 2021, pp. 14 816–14 826.

[42] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxim: Multi-axis mlp for image processing," in *Proc. CVPR*, 2022, pp. 5769–5780.

[43] J. Tian, J. Sun, and Y. Tang, "Tricolor attenuation model for shadow detection," *IEEE TIP*, vol. 18, no. 10, pp. 2355–2363, 2009.

[44] L. Qu, J. Tian, Z. Han, and Y. Tang, "Pixel-wise orthogonal decomposition for color illumination invariant and shadow-free image," *Optics Express*, vol. 23, no. 3, p. 2220, 2015.

[45] Y. Liu, Y. Li, S. You, and F. Lu, "Unsupervised learning for intrinsic image decomposition from a single image," in *Proc. CVPR*, 2020, pp. 3245–3254.

[46] P. Das, S. Karaoglu, and T. Gevers, "Pie-net: Photometric invariant edge guided network for intrinsic image decomposition," in *Proc. CVPR*, 2022, pp. 19 790–19 799.

[47] Y.-Y. Chuang, D. B. Goldman, B. Curless, D. H. Salesin, and R. Szeliski, "Shadow matting and compositing," *ACM TOG*, vol. 22, no. 3, pp. 494–500, 2003.

[48] 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.

[49] N. Inoue and T. Yamasaki, "Learning from synthetic shadows for shadow detection and removal," *IEEE TCSVT*, vol. 31, no. 11, pp. 4187–4197, 2020.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, Y. Bengio and Y. LeCun, Eds., 2015.

[51] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," *Proc. CVPR*, vol. 42, no. 8, pp. 7132–7141, 2018.

[52] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. CVPR*, 2015, pp. 447–456.

[53] X. Zhang, R. Ng, and Q. Chen, "Single Image Reflection Separation with Perceptual Losses," in *Proc. CVPR*, 2018, pp. 4786–4794.

[54] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," in *Proc. ICLR*, 2017.

[55] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proc. ECCV*, 2018, pp. 116–131.

[56] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv*, 2017.

[57] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. CVPR*, 2017, pp. 1251–1258.

[58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 448–456.

[59] Q. Chen, J. Xu, and V. Koltun, "Fast Image Processing with Fully-Convolutional Networks," in *Proc. ICCV*, vol. 2017-October, 2017, pp. 2516–2525.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE TPAMI*, vol. 37, no. 9, pp. 1904–1916, 2015.

[61] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, vol. 3, no. January, 2014, pp. 2672–2680.

[62] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017, pp. 5967–5976.

[63] 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.

[64] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *Proc. ICLR*, 2015.

[65] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proc. ICCV*, 2017, pp. 2242–2251.

[66] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *Proc. ICLR*, 2018.

[67] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in *Proc. ICCVW*, 2019, pp. 3265–3274.

[68] Q. Yang, K. H. Tan, and N. Ahuja, "Shadow removal using bilateral filtering," *IEEE TIP*, vol. 21, no. 10, pp. 4361–4368, 2012.

[69] R. Guo, Q. Dai, and D. Hoiem, "Paired regions for shadow detection and removal," *IEEE TPAMI*, vol. 35, no. 12, pp. 2956–2967, 2013.

[70] Z. Zou, S. Lei, T. Shi, Z. Shi, and J. Ye, "Deep Adversarial Decomposition: A Unified Framework for Separating Superimposed Images," in *Proc. CVPR*, 2020, pp. 12 803–12 813.

[71] H. Duan, W. Shen, X. Min, Y. Tian, J.-H. Jung, X. Yang, and G. Zhai, "Develop then rival: A human vision-inspired framework for superimposed image decomposition," *IEEE TMM*, 2022.

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

[73] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," in *Proc. CVPR*, 2018, pp. 6848–6856.

[74] X. Shu, J. Yang, R. Yan, and Y. Song, "Expansion-squeeze-excitation fusion network for elderly activity recognition," *TCSVT*, 2022.

[75] X. Shu, B. Xu, L. Zhang, and J. Tang, "Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition," *TPAMI*, 2022.

[76] B. Xu, X. Shu, and Y. Song, "X-invariant contrastive augmentation and representation learning for semi-supervised skeleton-based action recognition," *TIP*, 2022.

[77] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," in *Proc. CVPR*, 2020, pp. 13 906–13 915.