

A Multi-task Mean Teacher for Semi-supervised Shadow Detection

Zhihao Chen¹, Lei Zhu^{2,1}, Liang Wan^{1†}, Song Wang^{1,3}, Wei Feng¹, and Pheng-Ann Heng^{2,4}

¹ College of Intelligence and Computing, Tianjin University

² Department of Computer Science and Engineering, The Chinese University of Hong Kong

³ Department of Computer Science and Engineering, University of South Carolina

⁴ Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

Abstract

Existing shadow detection methods suffer from an intrinsic limitation in relying on limited labeled datasets, and they may produce poor results in some complicated situations. To boost the shadow detection performance, this paper presents a multi-task mean teacher model for semi-supervised shadow detection by leveraging unlabeled data and exploring the learning of multiple information of shadows simultaneously. To be specific, we first build a multi-task baseline model to simultaneously detect shadow regions, shadow edges, and shadow count by leveraging their complementary information and assign this baseline model to the student and teacher network. After that, we encourage the predictions of the three tasks from the student and teacher networks to be consistent for computing a consistency loss on unlabeled data, which is then added to the supervised loss on the labeled data from the predictions of the multi-task baseline model. Experimental results on three widely-used benchmark datasets show that our method consistently outperforms all the compared state-of-the-art methods, which verifies that the proposed network can effectively leverage additional unlabeled data to boost the shadow detection performance.

1. Introduction

As a common phenomenon in our daily life, shadows in natural images have hints for extracting the scene geometry [29, 17], light direction [22], camera location and its parameters [16], and benefit different high-level image understanding tasks, e.g., image segmentation [4], object detection [2], and object tracking [27]. For these applications, we need to detect shadows from images with high accuracy.

Existing methods detect shadows by developing physical models of color and illumination [6, 5], by using data-driven

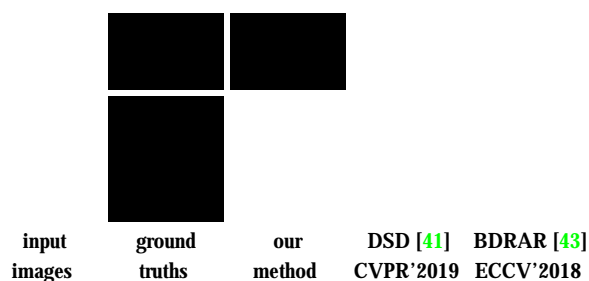


Figure 1: Shadow detection on two inputs with a soft shadow (the first row) and multiple shadow regions (the second row). Results in 3rd to 5-th columns are produced by our method, DSD [41], and BDRAR [43]. Apparently, our method can more accurately identify the shadow regions, while some dark regions, as well as shadow boundaries are mistakenly recognized by DSD and BDRAR.

approaches based on hand-crafted features [13, 23, 42] or by learning discriminative features from a convolutional neural network (CNN) [19, 33, 28, 12, 24, 43, 10, 41]. While the state-of-the-art methods have already achieved high accuracy on benchmark datasets [33, 42, 35, 10], they almost require sufficient amounts of annotated data for training, and such training data are usually captured in limited scenes. Creating large labeled datasets for diverse scenes, however, is expensive and time-consuming. Le et al. [24] proposed to augment training images by weakening the shadow area of the original training image, but we notice that these augmented images tend to be fake, and their non-shadow backgrounds are similar to those on the original training image, hindering the generalization capability. Compared with labeled datasets, we could easily collect abundant unlabeled shadowed images in real applications. Hence, it is highly desirable to leverage additional unlabeled data to improve the shadow detection performance when training with limited labeled data.

On the other hand, when testing the existing methods on various natural images, we found that they may ne-

Zhihao Chen and Lei Zhu are the joint first authors of this work.

[†]Corresponding author

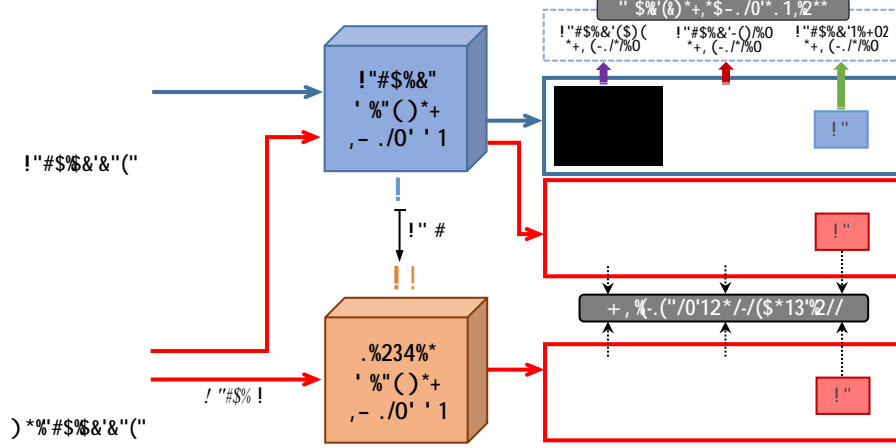


Figure 2: The schematic illustration of our multi-task mean teacher network (MTMT-Net). We first develop a multi-task CNN (MT-CNN; see Fig. 3) to mutually learn three tasks including shadow edge detection, shadow region detection, and shadow count detection. After that, we compute a multi-task supervised loss for labeled data and a multi-task consistency loss for unlabeled data. Finally, we fuse the supervised loss and consistency loss to train our shadow detection network.

glect small shadow regions, misrecognize dark regions as shadows, and miss non-obvious or soft shadows due to the weak boundaries. These situations result in poor shadow boundaries and may alter the number of shadow regions (see Fig. 1). Inspired by the success of multi-task learning in many computer vision applications [14, 3, 18, 26], we decide to investigate the complementary information of shadow regions, shadow edges and shadow count in our work, to enhance shadow region detection from both global and detail views. Specifically, shadow count detection sets a global constraint on the total number of shadow regions, while shadow edge detection sets detail-level constraints on the boundaries of shadow regions.

In this regard, we develop a multi-task mean teacher framework (MTMT-Net) for boosting the shadow detection performance. We first design a multi-task CNN, denoted as MT-CNN, for mutually learning three tasks (i.e., shadow region detection, shadow edge detection, and shadow count detection), and take this MT-CNN model as both the student network and the teacher network. We then propose a supervised multi-task loss for labeled data to integrate the supervised losses on all three tasks. After that, we enforce the three tasks' results of the student network and the teacher network to be consistent, respectively, on all the unlabeled data. By adding the supervised loss from the developed MT-CNN and the consistency loss from the three tasks to train the model, our network can more accurately detect shadow regions than the state-of-the-art methods. Our major contributions are summarized as:

- First, we develop a multi-task CNN (MT-CNN) for shadow detection by simultaneously detecting shadow regions, shadow edges, and shadow count from the single input image. The MT-CNN can produce a better shadow detection result on labeled data than the one

with only shadow detection task.

- Second, we propose to design a multi-task mean teacher framework to fuse consistency loss of unlabeled data from three prediction tasks for shadow detection. As a self-ensembling model, our framework has the potential to be used for developing semi-supervised frameworks on other vision tasks, including saliency detection, boundary detection, and semantic segmentation.
- Lastly, we show that the proposed network outperforms the state-of-the-art methods by a large margin on three widely-used benchmark datasets.

2. Related Work

Traditional methods. Early attempts [6, 5, 32] explored illumination models and color information to identify shadow regions and most of them work well only on high-quality and well-constrained images [28, 41]. Later data-driven strategies design certain hand-crafted features [42, 23, 7, 13, 34] on annotated data and feed these features into different classifiers [42, 23, 7, 13, 34] for shadow detection. Although achieving accuracy improvements, these strategies usually suffer from degraded performance in complex cases where hand-crafted features are not sufficiently discriminative for detecting shadow regions.

Deep learning based methods. Inspired by the remarkable progress of deep learning in diverse vision tasks, convolutional neural network (CNN) based methods have been developed for shadow detection to learn deep shadow inference features from labeled datasets. Khan et al. [19] formulated the first network to classify image pixels as shadows/non-shadows by building a 7-layer CNN, which extracts deep features from superpixels, and then feeding

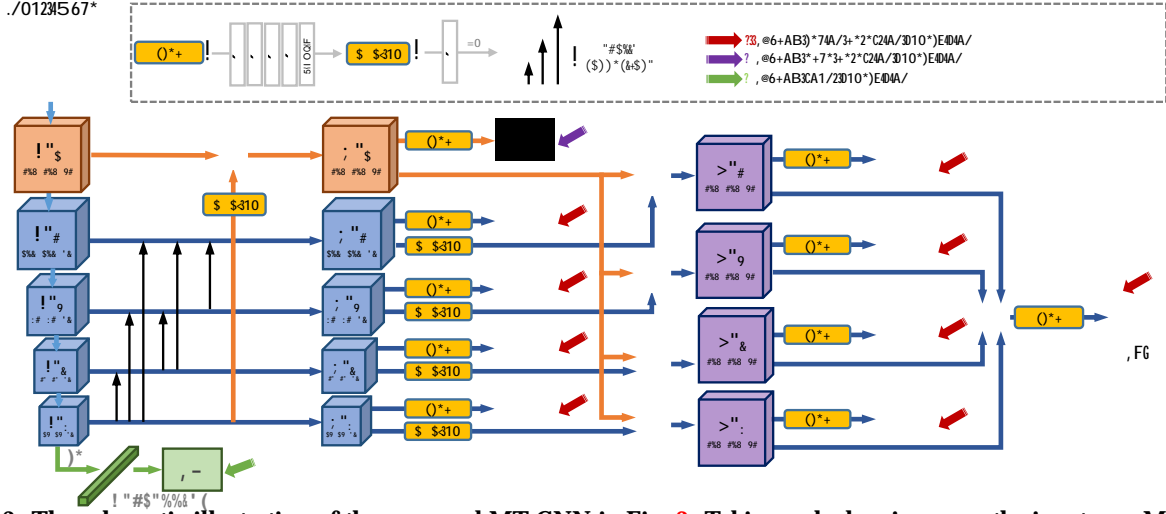


Figure 3: The schematic illustration of the proposed MT-CNN in Fig. 2. Taking a shadow image as the input, our MT-CNN predicts a shadow region map, a shadow edge map, and a shadow count (i.e., the number of shadow regions) by fusing their complementary information; see Section 3.1 for details.

the features to a conditional random field (CRF) model to smooth the shadow detection results. Vicente et al. [33] learned an image-level shadow prior and combined it with local image patches to train a patch-based CNN for generating a shadow mask. Later, a generative adversarial network based shadow detector, called scGAN [28], predicts a shadow map by formulating a conditional generator on the input image. A fast deep shadow detection network in [8] obtains a shadow prior map from hand-crafted features, applies a patch-level CNN to predict shadow masks of patches, and combines the results from multi-scale patches for predicting the whole shadow map.

Recently, Hu et al. [12] detected shadow pixels by learning direction-aware spatial context features. Zhu et al. [43] designed a recurrent attention residual (RAR) module to combine the contexts of two adjacent CNN layers and then formulated two series of RAR modules to iteratively integrate spatial contexts over the CNN layers. Le et al. [24] combined a shadow detection network (D-Net) with a shadow attenuation network (A-Net) that generated adversarial training examples. Wang et al. [37] stacked multiple parallel fusion branches to fuse global semantic cues and local spatial details in a deeply supervised framework. Zheng et al. [41] presented a distraction-aware shadow (DS) module to predict false positive and false negative pixels, and fused the obtained distraction features in each CNN layer for shadow detection.

Although improving the shadow detection bar, existing methods almost suffer from an intrinsic limitation that training their detection networks requires a large amount of data with pixel-level annotations. Although ADNet [24] augments training images from a single shadow image by weakening the shadow area, we argue that these augmented

images tend to be fake, and the backgrounds are very similar to the original training image, resulting in a limited generalization capability. In this paper, we leverage unlabelled data for helping shadow detection. For this purpose, we embed a multi-task learning into a self-ensembling framework to enforce consistency loss of shadow-detection tasks. Results show that our method outperforms state-of-the-art shadow detectors as detailed in the later experiment section.

3. Methodology

Fig. 2 shows the workflow of the proposed MTMT-Net that integrates labeled data and unlabeled data by using the mean teacher semi-supervised learning. Specifically, we develop a multi-task convolutional neural network (MT-CNN) by considering three tasks, i.e., shadow region detection, shadow edge detection, and shadow count detection. MT-CNN is used for both the student network and the teacher network. During the training, the labeled data is fed into the student network, and a multi-task supervised loss is computed by fusing the three task losses. Then, for unlabeled data, we produce one auxiliary shadow map from the input image and feed them into the student network and teacher network, respectively. A multi-task consistency loss is computed on the two groups of predicted shadow information. In the testing stage, we only utilize the student network to predict the shadow map for the input image.

3.1. Multi-task Convolutional Neural Network (MT-CNN)

Although achieving remarkable results, existing shadow detection methods suffer from performance degradation when detecting soft shadows due to the weak boundaries. They also tend to neglect small shadow regions or mis-

identify dark non-shadow regions, thereby may significantly alter the count of detected shadow regions. To address these concerns, we argue that explicitly considering shadow edges and shadow count is helpful to augment shadow region detection in both localization accuracy and segmentation quality. In this paper, we proposed a multi-task CNN (MT-CNN) to model and fuse the complementary of shadow edge, shadow count, and shadow region information within a single network in an end-to-end manner, as illustrated in Fig. 3.

3.1.1 Shadow Region Detection

Given an input shadow image, we first use a convolutional neural network (ResNeXt-101[38] in our experiment) to produce a set of feature maps (denoted as EF_1, EF_2, EF_3, EF_4 , and EF_5) at different scales (see Fig. 3).

Note that there is complementary information among different CNN layers for shadow detection. The shallow CNN layers capture shadow details as well as many non-shadow details, while the deep CNN layers neglect most of the non-shadow pixels and also miss parts of shadow regions. Here, we employ the short connections [9] to merge feature maps at the last four CNN layers, resulting in four new feature maps (denoted as DF_2, DF_3, DF_4 , and DF_5). Specifically, the merged feature map DF_k at k -th CNN layer ($k = 2, \dots, 5$) is computed by:

$$DF_k = \text{Conv}(\text{Concat}(EF_k, \dots, EF_5)). \quad (1)$$

We then merge the shallowest features (EF_1) and the deepest features (EF_5) to generate a new feature map, denoted as DF_1 , which is used for predicting shadow edge map (see Section 3.1.2). After that, to integrate the shadow edge and shadow region information, we refine $\{DF_k, k = 2, \dots, 5\}$ by first up-sampling them into the spatial resolution of DF_1 and then element-wise adding DF_1 . The refined feature maps are denoted as $\{RF_k, k = 2, \dots, 5\}$, given by

$$RF_k = \text{up}(DF_k) + DF_1. \quad (2)$$

Finally, we predict four shadow region maps from DF_2, DF_3, DF_4 , and DF_5 , four shadow region maps from RF_2, RF_3, RF_4 , and RF_5 , and a shadow map (denoted as S_f in Fig. 3) from the refined feature maps, which is produced by element-wisely adding, i.e.

$$S_f = \text{Pred}\left(\sum_{k=2}^5 RF_k\right). \quad (3)$$

The prediction $\text{Pred}(\cdot)$ is realized by using three 3×3 convolutional layers, a 1×1 convolutional layer, and a sigmoid activation layer [43] on features.

3.1.2 Shadow Edge Detection

By observing shadow images, we notice that for soft shadows, the boundaries may not be distinguishable from the

surrounding non-shadow regions. This motivates us to think about utilizing edge knowledge to enhance the detection performance. Recent saliency detectors [14, 15] have also proved this point, in which edge knowledge is helpful to improve the saliency detection quality.

In our MT-CNN, we fuse the low-level CNN features EF_1 with the high-level features EF_5 at the deepest CNN layer to produce the feature map DF_1 , which is then used for predicting a shadow edge map. Although low-level features EF_1 capture sufficient shadow edge information, detecting shadow edges only with EF_1 is not sufficient, since EF_1 also encodes many non-shadow background details. On the other hand, the deep layer features EF_5 has the largest receptive field to effectively suppress the non-shadow pixels. Specifically, DF_1 is computed via an element-wise addition on EF_1 and EF_5 .

3.1.3 Shadow Count Detection

By analyzing the results of existing shadow detection methods, we find three common failure cases: small shadow regions are missed; non-shadow regions are mis-identified; and nearby shadow regions are mistakenly detected together. These cases all result in an inaccurate shadow region number. Therefore, we explore the number of shadow regions for enhancing the shadow detection performance.

Detecting the shadow region number requires a global understanding of the whole image. As shown in Fig. 3, we rely on EF_5 at the deepest CNN layer for the detection. Specifically, we apply a single fully-connected layer on EF_5 to obtain a score (A) indicating the shadow count. Since the number of shadow regions can be very large, to make the computation feasible, we set a maximum constraint N_{\max} , and empirically compute the scalar A as the regression problem:

$$A = \frac{\min(N_{\text{actual}}, N_{\max})}{N_{\max}}, \quad (4)$$

where N_{actual} denotes the actual number of the shadow regions, and we empirically set $N_{\max}=8$ in our work.

3.2. Multi-task Supervised Loss on Labeled data

For labeled data, we can have a pair of input shadow image and the corresponding annotated shadow mask. It is natural that we take the annotated shadow mask as the ground truth of the shadow region detection (G_r). Then, we apply the Canny operator [1] on the annotated shadow mask to generate an edge map as the ground-truth of the shadow edge detection (G_e). We further observe each labeled image and manually count the number of shadow regions to obtain A (see Eq. (4)), which is regarded as the ground truth of the shadow count detection (G_c).

After obtaining the ground truths, the multi-task supervised loss (denoted as L^s) for a labeled image (x) is computed by adding the supervised losses of the shadow region

detection (L_r^s), shadow edge detection (L_e^s), and shadow count detection (L_c^s), i.e.

$$L^s(x) = L_r^s + L_e^s + L_c^s, \quad (5)$$

where

$$\begin{aligned} L_r^s &= \sum_{j=1}^9 \text{BCE}(P_r(j), G_r), \\ L_e^s &= \text{BCE}(P_e, G_e), \\ L_c^s &= \text{MSE}(P_c, G_c). \end{aligned} \quad (6)$$

Here, $P_r(j)$ represents one of the nine predicted shadow maps, P_e is the predicted shadow edge map, and P_c is the predicted shadow count value. BCE and MSE are the binary cross-entropy loss and MAE loss functions, respectively. We empirically set the weights $\alpha=10$ and $\beta=1$ in the network training.

3.3. Multi-task Consistency Loss on Unlabeled Data

For the unlabeled data, we pass it into the student and teacher networks to obtain three tasks' results, which are the nine shadow region maps (denoted as S_{r_1} to S_{r_9}), a shadow edge map (denoted as S_e), and a shadow count score (denoted as S_c). We then enforce the predictions of the three tasks from the student network and teacher network to be consistent, resulting in a multi-task consistency loss (L^c). Mathematically, L^c for an unlabeled image (denoted as y) is

$$L^c(y) = L_r^c + L_e^c + L_c^c \quad (7)$$

where

$$\begin{aligned} L_r^c &= \sum_{j=1}^9 \text{MSE}(S_{r_j}, T_{r_j}), \\ L_e^c &= \text{MSE}(S_e, T_e), \\ L_c^c &= \text{MSE}(S_c, T_c), \end{aligned} \quad (8)$$

where L_r^c , L_e^c , and L_c^c denote the consistency loss of the shadow region detection, shadow edge detection, and shadow count detection, respectively.

3.4. Our Network

We apply the multi-task learning with the semi-supervised self-ensembling model for shadow detection. The total loss of our network is

$$L_{\text{total}} = \sum_{i=1}^N L^s(x_i) + \sum_{j=1}^M L^c(y_j), \quad (9)$$

where N and M are the numbers of labeled images and unlabeled images in our training set. $L^s(x_i)$ denotes the multi-task supervised loss (Eq. (5)) for the i -th labeled image while $L^c(y_j)$ is the multi-task consistency loss (Eq. (7)) for the j -th unlabeled image. The weight α is to balance the multi-task supervised loss on labeled data and the multi-task consistency loss on unlabeled data. Following [21, 31], we

use a time dependent Gaussian warming up function to update $\alpha(t) = \max e^{(-5(1-t/t_{\max})^2)}$, where t denotes the current training iteration and t_{\max} is the maximum training iteration. In our experiments, we empirically set $t_{\max}=10$.

We minimize L_{total} to train the student network, and the parameters of the teacher network in each training step, are updated via the exponential moving average (EMA) strategy in [31]. The parameters of the teacher network at the t training iteration are:

$$\theta_t = \alpha \theta_{t-1} + (1 - \alpha) \theta_t, \quad (10)$$

where θ_t denotes the student network parameter at the t training iteration. The EMA decay α is empirically set as 0.99, as suggested in [21, 31].

Our unlabeled data. The unlabeled data in our work has 3,424 images with shadows. It consists of two parts: one is the USR dataset from a recent shadow removal work [11], while the other is our collection of 979 images from the internet. The USR dataset [11] has 2,445 shadow images without shadow detection annotations.

3.5. Training and Testing Strategies

Training parameters. To accelerate the training procedure and reduce the overfitting risk, we initialize the parameters of MT-CNN (student network) by ResNeXt [38], which has been well-trained for the image classification task on the ImageNet. Other parameters in the MT-CNN are initialized as random values. Stochastic gradient descent (SGD) equipped with a momentum of 0.9 and a weight decay of 0.0005 is used to optimize the whole network with 10,000 iterations. The learning rate is adjusted by a poly strategy [25] with the initial learning rate of 0.005 and the power of 0.9. We resize all the labeled and unlabeled images to 416×416 for training our network on a single GTX 2080Ti GPU, and augment the training set by random horizontal flipping. We use the mini-batch size of 6, which consists of 4 labeled images and 2 unlabeled data images.

Inference. During testing, we resize the input images to 416×416 , feed the resized image into the student network, and take the rightmost shadow region detection map (S_f in Fig. 3) as the final output of our MTMT-Net. Following recent shadow detection networks [43, 41], we apply a fully connected conditional random field (CRF) [20] to further post-process the predicted result of our network.

4. Experimental Results

In this section, we first present the shadow detection benchmark datasets and evaluation metric, then compare the proposed MTMT-Net with the state-of-the-art shadow detectors and those relevant works including shadow removal, saliency detection and semantic segmentation, and

Table 1: Comparing our network (MTMT-Net) against the state-of-the-art shadow detectors.

Method	Year	SBU [33]			UCF [42]			ISTD [35]		
		BER	Shadow	Non Shad.	BER	Shadow	Non Shad.	BER	Shadow	Non Shad.
MTMT-Net(ours)	-	3.15	3.73	2.57	7.47	10.31	4.63	1.72	1.36	2.08
Ours w/o-CRF	-	3.15	3.72	2.58	8.06	12.23	3.90	1.77	1.16	2.39
DSDNet [41]	2019	3.45	3.33	3.58	7.59	9.74	5.44	2.17	1.36	2.98
DC-DSPF [37]	2019	4.90	4.70	5.10	7.90	6.50	9.30	-	-	-
BDRAR [43]	2018	3.64	3.40	3.89	7.81	9.69	5.94	2.69	0.50	4.87
ADNet [24]	2018	5.37	4.45	6.30	9.25	8.37	10.14	-	-	-
DSC [12]	2018	5.59	9.76	1.42	10.54	18.08	3.00	3.42	3.85	3.00
ST-CGAN [35]	2018	8.14	3.75	12.53	11.23	4.94	17.52	3.85	2.14	5.55
patched-CNN [8]	2018	11.56	15.60	7.52	-	-	-	-	-	-
scGAN [28]	2017	9.10	8.39	9.69	11.50	7.74	15.30	4.70	3.22	6.18
stacked-CNN [33]	2016	11.00	8.84	12.76	13.00	9.00	17.10	8.60	7.69	9.23
Unary-Pairwise [7]	2011	25.03	36.26	13.80	-	-	-	-	-	-
DeshadowNet [30]	2017	6.96	-	-	8.92	-	-	-	-	-
EGNet [14]	2019	4.49	5.23	3.75	9.20	11.28	7.12	1.85	1.75	1.95
SRM [36]	2017	6.51	10.52	2.50	12.51	21.41	3.60	7.92	13.97	1.86
Amulet [39]	2017	15.13	-	-	15.17	-	-	-	-	-
PSPNet [40]	2017	8.57	-	-	11.75	-	-	4.26	4.51	4.02

finally report ablation study results. Our code, model parameters, and shadow detection results on three benchmark datasets have been released at <https://github.com/eraserNut/MTMT>.

4.1. Datasets and Evaluation Metrics

Benchmark datasets. We evaluate our method on three widely-used shadow detection benchmark datasets: SBU [33], UCF [42], and ISTD [35]: (i) The SBU dataset is the largest annotated shadow dataset with 4,089 training images and 638 testing images; (ii) The UCF dataset consists of 145 training images and 76 testing images, covering outdoor scenes; and (iii) ISTD is a recently developed dataset for both shadow detection and removal. It has 1,870 triples of shadow images, shadow maps, and shadow-free images, and 540 of them are used for testing. Similar to recent works [12, 24, 43, 41], for SBU and UCF, we obtained the evaluation results by training our network on the SBU training set and our unlabeled dataset. Since ISTD only contains cast shadow images that are different from SBU images, following [41], we re-train our method and most competitors on the ISTD training dataset with our unlabeled data. Our training time for SBU is 1 hour, and that for ISTD is 0.5 hours. The model size of our network is 169 M. In the testing, our MTMT-Net takes around 0.05 seconds to process an image with a 416×416 image resolution.

Evaluation metric. We employ a commonly-used metric, i.e., balance error rate (BER), to quantitatively evaluate the shadow detection performance. The BER [43, 12] equally considers the quality of shadow and non-shadow regions, which is given by:

$$\text{BER} = 1 - \frac{1}{2} \left(\frac{N_{tp}}{N_p} + \frac{N_{tn}}{N_n} \right) \times 100, \quad (11)$$

where N_{tp} , N_{tn} , N_p , and N_n are the number of true positives, true negatives, shadow and non-shadow pixels of the

shadow image, respectively. A small BER value indicates a better shadow detection performance.

4.2. Comparison with the State-of-the-art Shadow Detectors

We make comparison with 10 recent shadow detectors including DSDNet [41], DC-DSPF [37], BDRAR [43], ADNet [24], DSC [12], ST-CGAN [35], patched-CNN [8], scGAN [28], stacked-CNN [33], and Unary-Pairwise [7]. Among them, the last method is based on hand-crafted features while all the others are deep-learning-based methods. To make the comparisons fair, we adopt the available results of compared methods by either directly from the authors or using their report in published paper.

Quantitative comparison. Table 1 summarizes the quantitative results of different methods on the three benchmark datasets. The BER score is the average of shadow and non-shadow BER scores. Apparently, the deep learning based methods [33, 12, 8] have much smaller BER values than the hand-crafted detector [7], since they can learn more powerful features for shadow detection from the annotated training images. Among the deep learning based shadow detectors, DSDNet [41] is the second best-performing method, which explicitly learns and integrates the semantics of visual distraction regions to infer shadows. Compared to the best-performing existing method, our method has 8.70%, 1.58%, and 20.7% lower BER scores on SBU, UCF, and ISTD, respectively. In addition, our method has a better BER score on non-shadow pixels for SBU and UCF and a better BER score on shadow pixels for ISTD. This shows that our network predicts more shadow pixels for SBU and UCF and reduces the false positive predictions on non-shadow regions for ISTD. Like the three comparative methods [43, 12, 41], we also use CRF [20] as post-processing. The second row in Table 1 shows the performance of our

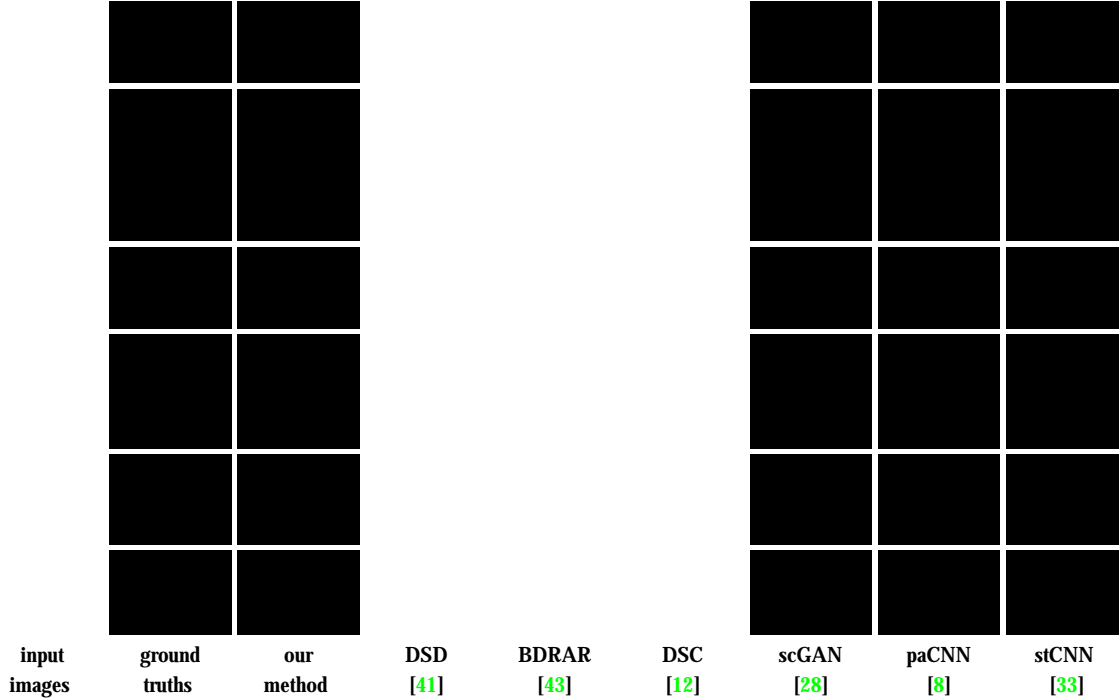


Figure 4: Visual comparison of shadow maps produced by our method and other methods (4th-10th columns) against ground truths shown in 2nd column. Note that “stCNN” and “paCNN” stand for “stacked-CNN” and “patched-CNN”, respectively.

method without using CRF. The results indicate using CRF obtains a certain degree of improvement, mainly on the UCF dataset, while without CRF still achieves better performance than most state-of-the-art methods.

Visual comparison. We further visually compare the shadow detection maps produced by our method and the state-of-the-arts, as shown in Figs. 4. From the results, we can see that our MTMT-Net (3rd column of Figs. 4) has the best performance among all the shadow detectors. It can effectively locate different shadows under various backgrounds, and successfully discriminate true shadows from those non-shadow regions with shadow appearance. For example, in the 3rd, 5th and 7th row, MTMT-Net can accurately detect the shadow regions, while the others mistakenly recognize the road, the sky and the dark ground as shadows, respectively. What’s more, for high-contrast objects in a large shadow region, MTMT-Net can still recognize them as shadows, as demonstrated in the last two rows.

4.3. Comparison with Shadow Removal, Saliency Detection and Semantic Segmentation Methods

It is noted that deep networks designed for shadow removal, saliency detection and semantic segmentation can be re-trained for shadow detection by using annotated shadow datasets. To further evaluate the effectiveness of our method, we apply a shadow removal model, i.e., DeshadownNet [30], three saliency detection models, i.e., SRM [36],

Amulet [39] and EGNet [14], and a semantic segmentation model, i.e., PSPNet [40] on shadow detection datasets.

We adopt the available results of compared methods by either re-training the released code or using those reported. For a fair comparison, we try our best to fine tune their training parameters and select the best shadow detection results. The last five rows in Table 1 report their BER values. We see that these models can achieve superior BER performance over some existing shadow detectors, yet are still worse than our network.

4.4. Ablation Analysis

Baseline network design. We perform ablation study experiments to evaluate the proposed multi-task supervised loss (see Eq. (5)) and multi-task consistency loss (see Eq. (7)) of our MTMT-Net. Here, we consider seven baseline networks.

The first four baseline networks are constructed by removing the teacher model. It means that only supervised loss on labeled data is used to train MT-CNN. Specifically, the first baseline network (denoted as “basic”) only considers the shadow region detection supervised loss (L_r^s in Eq. (5)). The second (denoted as “basic+SE”) is to add the shadow edge detection supervised loss (L_e^s of Eq. (5)), while the third (denoted as “basic+SC”) is to add shadow count detection supervised loss (L_c^s of Eq. (5)). The fourth is to fuse the supervised loss of three tasks together.

Another three baseline networks are built to verify the multi-task consistency loss on unlabeled data. The first one

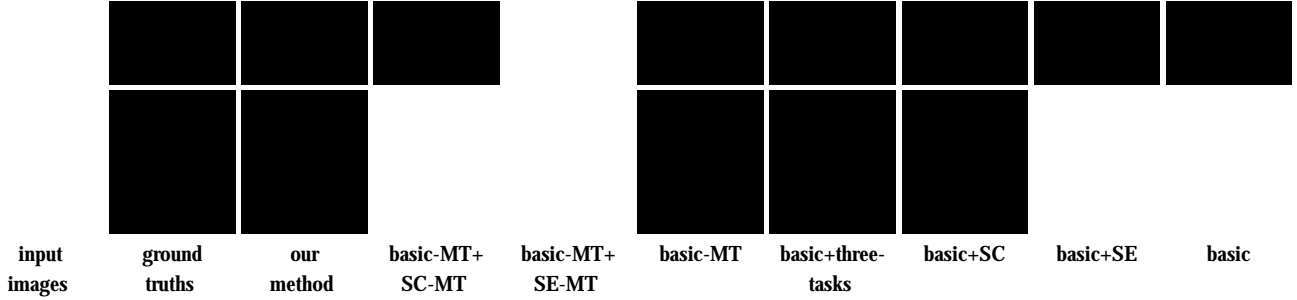


Figure 5: Visual comparison of shadow maps produced by our method and other baseline networks (see Table 2 for details).

Table 2: Ablation analysis. Here, “SR” denotes the shadow region detection; “SE” denotes the shadow edge detection; “SC” denotes the shadow count detection; and “MT” denotes the mean teacher.

Network	SR	SE	SC	MT	SBU [33] BER	UCF [42] BER	ISTD [35] BER
basic		×	×	×	5.28	9.57	2.23
basic+SE			×	×	4.07	8.09	1.8
basic+SC		×		×	4.72	9.34	2.04
basic+three-tasks				×	3.61	7.64	2.03
basic-MT		×	×		4.49	8.29	2.12
basic-MT+SE-MT			×		3.83	7.81	1.75
basic-MT+SC-MT		×			4.41	8.61	2.03
our method					3.15	7.34	1.72

(denoted as “basic-MT”) only considers the mean teacher model on the shadow region task by fusing the supervised loss (L_r^s of Eq. (5)) and the consistency loss (L_r^c of Eq. (7)). The second one (denoted as “basic-MT+SE-MT”) is to apply the mean teacher model on the shadow region detection and shadow edge detection, which means that L_r^s and L_e^s in Eq. (5) as well as L_r^c and L_e^c in Eq. (7) are used to train the network. The last one (denoted as “basic-MT+SC-MT”) is to use the mean teacher model on the shadow region detection and shadow count detection. In other words, the supervised loss (L_r^s and L_c^s in Eq. (5)) and the consistency loss (L_r^c and L_c^c in Eq. (7)) are used for the network training.

We train all the seven baseline networks using the SBU training set and our unlabeled data to obtain results of SBU and UCF. For ISTD, we use the ISTD training set and our unlabeled data to train all four networks and test them using the ISTD testing set.

Quantitative comparisons. Table 2 summaries the BER values of our network and seven baseline networks on the three benchmark datasets. From the results, we have the following observations: (i) “basic+SE” and “basic+SC” have superior BER values over “basic”, which means that detecting shadow boundaries and shadow count can provide helpful information for shadow detection. (ii) “basic+three-tasks” has better BER performance than “basic+SE” and “basic+SC”, demonstrating that fusing the three tasks for a supervised shadow detection together incurs a better shadow detection performance. (iii) “basic-MT” can more accurately detect shadow pixels than “basic” due to its smaller BER values. It indicates that the additional con-

sistency loss from the unlabeled data incurs a superior shadow detection performance. (iv) “basic-MT+SE-MT” and “basic-MT+SC-MT” produce smaller BER results than “basic-MT”, showing that the shadow edge detection and shadow region detection benefit the mean teacher model for shadow detection. (v) We can find that the shadow edge detection has a more contribution than the shadow count detection to the success of our method since “basic-MT+SE-MT” has a better BER result than “basic-MT+SC-MT”. (vi) By designing a three-task mean teacher model, our MTMT-Net has the best BER performance on three benchmarks.

Visual comparisons. Moreover, Fig. (5) visually compares shadow maps produced by our MTMT-Net and seven baseline networks. Apparently, our method can identify shadows better than all seven baselines in both shadow segmentation quality and localization accuracy. This proves the effectiveness of considering shadow edge, shadow count information and unlabeled data within one framework.

5. Conclusion

This paper presents a novel network for single-image shadow detection by developing a multi-task mean teacher framework. Our key idea is to first develop a multi-task network for simultaneously predicting shadow region detection, shadow edge detection, as well as shadow count estimation by leveraging their complementary information. Then we employ the mean teacher semi-supervised learning to leverage additional unlabeled data for further improving the detection performance. Experimental results on three benchmark datasets show that our network consistently outperforms the state-of-the-art methods by a large margin. Like other works [43, 41, 12], our method might not work well for images with multiple and complex shadows. Resolving this challenging problem is considered as a future direction of our work.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (Project No. 61572354, 61902275, 61671325, U1803264, 61672376), and CUHK Research Committee Direct Grant for Research 2018/19.

References

- [1] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986. [4](#)
- [2] Rita Cucchiara, Costantino Grana, Massimo Piccardi, and Andrea Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, 2003. [1](#)
- [3] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *ICCV*, pages 2051–2060, 2017. [2](#)
- [4] Aleksandrs Ecins, Cornelia Fermuller, and Yiannis Aloimonos. Shadow free segmentation in still images using local density measure. In *ICCV*, pages 1–8, 2014. [1](#)
- [5] Graham D Finlayson, Mark S Drew, and Cheng Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85(1):35–57, 2009. [1](#), [2](#)
- [6] Graham D Finlayson, Steven D Hordley, Cheng Lu, and Mark S Drew. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):59–68, 2006. [1](#), [2](#)
- [7] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Single-image shadow detection and removal using paired regions. In *CVPR*, pages 2033–2040, 2011. [2](#), [6](#)
- [8] Sepideh Hosseinzadeh, Moein Shakeri, and Hong Zhang. Fast shadow detection from a single image using a patched convolutional neural network. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3124–3129, 2018. [3](#), [6](#), [7](#)
- [9] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):815–828, 2019. [4](#)
- [10] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. to appear. [1](#)
- [11] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-ShadowGAN: Learning to remove shadows from unpaired data. In *ICCV*, pages 2472–2481, 2019. [5](#)
- [12] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *CVPR*, pages 7454–7462, 2018. [1](#), [3](#), [6](#), [7](#), [8](#)
- [13] Xiang Huang, Gang Hua, Jack Tumblin, and Lance Williams. What characterizes a shadow boundary under the sun and sky? In *ICCV*, pages 898–905, 2011. [1](#), [2](#)
- [14] Deng-Ping Fan Yang Cao Ju-Feng Yang Ming-Ming Cheng Jia-Xing Zhao, Jiang-Jiang Liu. EGNet: Edge guidance network for salient object detection. In *ICCV*, pages 8779–8788, 2019. [2](#), [4](#), [6](#), [7](#)
- [15] Ming-Ming Cheng Jiashi Feng Jianmin Jiang Jiang-Jiang Liu, Qibin Hou. A simple pooling-based design for real-time salient object detection. In *CVPR*, pages 3917–3926, 2019. [4](#)
- [16] Imran N Junejo and Hassan Foroosh. Estimating geotemporal location of stationary cameras using shadow trajectories. In *ECCV*, pages 318–331, 2008. [1](#)
- [17] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Trans. on Graphics (SIGGRAPH Asia)*, 30(6):157:1–157:12, 2011. [1](#)
- [18] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pages 7482–7491, 2018. [2](#)
- [19] Salman Hameed Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Automatic feature learning for robust shadow detection. In *CVPR*, pages 1939–1946, 2014. [1](#), [2](#)
- [20] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, pages 109–117, 2011. [5](#), [6](#)
- [21] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. [5](#)
- [22] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Estimating natural illumination from a single outdoor image. In *ICCV*, pages 183–190, 2009. [1](#)
- [23] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Detecting ground shadows in outdoor consumer photographs. In *ECCV*, pages 322–335, 2010. [1](#), [2](#)
- [24] Hieu Le, Yago Vicente, F Tomas, Vu Nguyen, Minh Hoai, and Dimitris Samaras. A+ D Net: Training a shadow detector with adversarial shadow attenuation. In *ECCV*, pages 662–678, 2018. [1](#), [3](#), [6](#)
- [25] Wei Liu, Andrew Rabinovich, and Alexander C Berg. ParseNet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015. [5](#)
- [26] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, pages 3994–4003, 2016. [2](#)
- [27] Sohail Nadimi and Bir Bhanu. Physical models for moving shadow and object detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):1079–1087, 2004. [1](#)
- [28] Vu Nguyen, Tomas F Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *ICCV*, pages 4510–4518, 2017. [1](#), [2](#), [3](#), [6](#), [7](#)
- [29] Takahiro Okabe, Imari Sato, and Yoichi Sato. Attached shadow coding: Estimating surface normals from shadows under unknown reflectance and lighting conditions. In *ICCV*, pages 1693–1700, 2009. [1](#)
- [30] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. DeshadowNet: A multi-context embedding deep network for shadow removal. In *CVPR*, pages 4067–4075, 2017. [6](#), [7](#)
- [31] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. pages 1195–1204, 2017. [5](#)

- [32] Jiandong Tian, Xiaojun Qi, Liangqiong Qu, and Yandong Tang. New spectrum ratio properties and features for shadow detection. *Pattern Recognition*, 51:85–96, 2016. 2
- [33] Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *ECCV*, pages 816–832, 2016. 1, 3, 6, 7, 8
- [34] Yago Vicente, F Tomas, Minh Hoai, and Dimitris Samaras. Leave-one-out kernel optimization for shadow detection. In *ICCV*, pages 3388–3396, 2015. 2
- [35] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, pages 1788–1797, 2018. 1, 6, 8
- [36] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, pages 4019–4028, 2017. 6, 7
- [37] Yupei Wang, Xin Zhao, Yin Li, Xuecai Hu, Kaiqi Huang, et al. Densely cascaded shadow detection network via deeply supervised parallel fusion. In *IJCAI*, pages 1007–1013, 2019. 3, 6
- [38] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 5987–5995, 2017. 4, 5
- [39] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *CVPR*, pages 202–211, 2017. 6, 7
- [40] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 6, 7
- [41] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. Distraction-aware shadow detection. In *CVPR*, pages 5167–5176, 2019. 1, 2, 3, 5, 6, 7, 8
- [42] Jiejie Zhu, Kegan GG Samuel, Syed Z Masood, and Marshall F Tappen. Learning to recognize shadows in monochromatic natural images. In *CVPR*, pages 223–230, 2010. 1, 2, 6, 8
- [43] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*, pages 121–136, 2018. 1, 3, 4, 5, 6, 7, 8