Densely Cascaded Shadow Detection Network via Deeply Supervised Parallel Fusion

Yupei Wang[1,2], **Xin Zhao**[1,2], **Yin Li**[3], **Xuecai Hu**[1,4], **Kaiqi Huang**[1,2,5]
[1] CRIPAC, NLPR, Institute of Automation, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] Carnegie Mellon University
[4] University of Science and Technology of China
[5] CAS Center for Excellence in Brain Science and Intelligence Technology
wangyupei2014@ia.ac.cn, {xzhao,kqhuang}@nlpr.ia.ac.cn,
yinl2@andrew.cmu.edu, xuecai.hu@cripac.ia.ac.cn

Abstract

Shadow detection is an important and challenging problem in computer vision. Recently, single image shadow detection had achieved major progress with the development of deep convolutional networks. However, existing methods are still vulnerable to background clutters, and often fail to capture the global context of an input image. These global contextual and semantic cues are essential for accurately localizing the shadow regions. Moreover, rich spatial details are required to segment shadow regions with precise shape. To this end, this paper presents a novel model characterized by a deeply supervised parallel fusion (DSPF) network and a densely cascaded learning scheme. The DSPF network achieves a comprehensive fusion of global semantic cues and local spatial details by multiple stacked parallel fusion branches, which are learned in a deeply supervised manner. Moreover, the densely cascaded learning scheme is employed to refine the spatial details. Our method is evaluated on two widely used shadow detection benchmarks. Experimental results show that our method outperforms state-of-the-arts by a large margin.

1 Introduction

Shadow occurs frequently in natural scenes. Patterns of shadows provide important cues for estimating physical properties of the scene e.g., light source [Lalonde *et al.*, 2010], illumination conditions [Panagopoulos *et al.*, 2012] and scene geometry [Karsch *et al.*, 2011]. Understanding regions of shadows can also help to improve downstream vision tasks, such as image segmentation [Ecins *et al.*, 2014]. Moreover, removing shadows from images has many applications in computational photography [Guo *et al.*, 2011; Qu *et al.*, 2017]. Among these tasks, shadow detection–identify regions of shadows, is a core step. We focus on the task of shadow detection from a single image in this paper.

Finding shadows from a single image is very challenging. The problem is fundamentally ill-posed, as scene geometry, surface property and lighting conditions can not be recovered from a 2D image. Early works made use of image prior and physical models of illumination and colors [Finlayson *et al.*, 2009; 2006]. These model based methods will fail if the underlie assumptions, such as Lambertian reflectance, are violated in natural images. To address this challenge, several recent works proposed learning-based approaches [Guo *et al.*, 2011; Zhu *et al.*, 2011; Tumblin and Williams, 2011; Khan *et al.*, 2016; Vicente *et al.*, 2018] for shadow detection. More recently, deep Convolutional Neural Networks (CNN) [Krizhevsky *et al.*, 2012] have been exploited, and significantly advanced the performance [Vicente *et al.*, 2016; Nguyen *et al.*, 2017] on public benchmarks.

Our method follows the same paradigm of learning deep models for shadow detection, and forms the problem as a dense binary labeling of pixels. Our work is different from [Vicente *et al.*, 2016; Nguyen *et al.*, 2017] by revisiting the idea of a fully convolutional network [Shelhamer *et al.*, 2017; Xie and Tu, 2015]. Specifically, we argue that both global context and local appearance are encoded within the hierarchy of a CNN. And we propose a novel architecture for combining these cues for shadow detection. Unlike [Vicente *et al.*, 2016], we use a single network and do not use separate models for global image prior and local patch appearance. Unlike [Nguyen *et al.*, 2017], we show that a binary entropy loss can outperform the adversarial loss with proper design of the network architecture. Fig 1 presents our results.

To this end, we propose a novel Densely Cascaded Deeply Supervised Parallel Fusion (DC-DSPF) network for shadow detection. Our DC-DSPF model equips an existing network [Xie and Tu, 2015] with two key components: (1) the deeply supervised parallel fusion; and (2) the densely cascaded learning. These two components are combined to obtain a better fusion of global context and local appearance in input

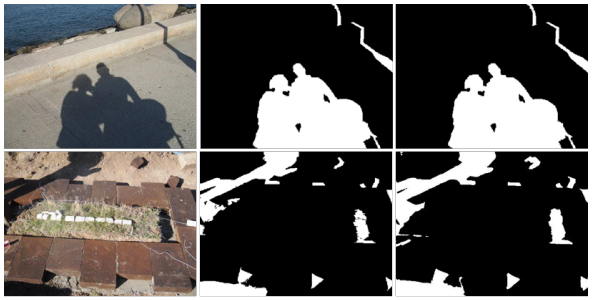

Figure 1: Results of shadow detection from our proposed DC-DSPF network. Our method can capture the fine details of shadow regions

images. The deeply supervised parallel fusion network seeks to combine both multi-scale visual features across the hierarchy of the backbone network. The densely cascaded learning scheme aims at integrating global contextual cues.

Specifically, our proposed Deeply Supervised Parallel Fusion (DSPF) stacks multiple parallel fusion units [Pinheiro *et al.*, 2016; Wang *et al.*, 2017b] on the backbone network. Each unit is a backward fusion branch that progressively merges feature maps from a previous unit (with the first one from the backbone network), and generates a full resolution output at the end. Loss functions are attached to all ends. By stacking multiple fusion branches together, DSPF network outputs multi-level feature maps, and learns to best fuse them in a deeply supervised manner [Lee *et al.*, 2015].

Moreover, our densely cascaded learning scheme sticks two DSPF networks together. We concatenate the initial prediction map of the first DSPF network and the original input image as the input of the second one. We also connect the multiple intermediate predictions generated by the parallel fusion branches in the first DSPF network to the corresponding intermediate predictions in the second one [Huang *et al.*, 2017]. In this case, multi-scale features in the previous DSPF network can be directly propagated to the next network. Therefore, the densely cascaded DSPF network can learn to better reason about the spatial context.

To verify our model, we conduct extensive experiments on the standard benchmarks. Our method outperforms the state-of-the-arts by a significant margin on the commonly used SBU [Vicente *et al.*, 2016] and UCF [Zhu *et al.*, 2011] datasets.

2 Densely Cascaded Network via Deeply Supervised Parallel Fusion

This section presents our Densely Cascaded Deeply Supervised Parallel Fusion (DC-DSPF) model for shadow detection. We start by introducing HED which is the basis of our model. Then we describe the key components of our proposed DC-DSPF model.

2.1 HED for Shadow Detection

Our proposed network is derived from the HED [Xie and Tu, 2015]. Thus we start by introducing the HED network. Note that we re-purpose HED for shadow region segmentation.

HED employs a deep CNN with multiple side outputs and generates per-pixel binary predictions. Specifically, HED network computes multi-scale prediction maps at all side-outputs and fuses these intermediate maps as the final output. HED thus leverages multi-scale representations given by the hierarchy of the network, and has been proven effective for detecting image boundaries.

Formally, we denote the training samples as $\{(X_n, Y_n)\}$, where $n = 1, ..., N$ indexes the samples. X_n is an input image and $Y_n = y_j^{(n)}, j = 1, ..., |X_n|, y_j^n \in \{0, 1\}$ is its ground truth shadow labels with pixel index j. We may drop the subscript n when it is clear from the context. HED is fully convolutional. And we denote the union of all convolutional weights in base network as W. HED further attaches M side-outputs to the network at each scale m. And we denote their parameters as $w = (w^{(1)}, ..., w^{(M)})$. $w^{(m)}$ can be considered as an individual linear classifier for features at scale m. For each scale, HED computes $P(y_j = 1|X; W, w^m) = \sigma(\hat{a}_j^{(m)})$ on its activation map $\hat{A}^{(m)}$ at all pixels with sigmoid function $\sigma(.)$. Their outputs $\hat{Y}^{(m)} = \sigma(\hat{A}^{(m)})$ forms the multi-scale outputs for shadow detection.

Fusion: These multi-scale side-output predictions are further linearly combined to generate the output prediction, as

$$\hat{Y} = \sigma(\hat{A}) = \sigma(\Sigma_{m=1}^{M} h_m \hat{A}^{(m)}) \quad (1)$$

with the fuse loss $L_{fuse} = l(Y, \hat{Y})$ as the loss between Y and $\hat{Y}$. Therefore, the final loss function of HED is given by

$$L_{HED}(W, w, h) = L_{side}(W, w) + L_{fuse}(W, w, h) \quad (2)$$

where $L_{side}(W, w)$ is the sum of all side-output loses, given by $L_{side}(W, w) = \Sigma_{m=1}^{M} l^{(m)}(W, w^{(m)})$.

Loss Function: HED uses weighted cross-entropy loss for each side-output, as a remedy to unbalanced samples. The loss function is given by,

$$\begin{aligned} l^{(m)}(W, w^{(m)}) = &- \beta \Sigma_{j \in Y_+} \log P(y_j = 1|X; W, w^{(m)}) \\ &- (1 - \beta) \Sigma_{j \in Y_-} \log P(y_j = 0|X; W, w^{(m)}), \end{aligned}$$

where $\beta = |Y_-|/|Y|$ and $1 - \beta = |Y_+|/|Y|$. Y_+ and Y_- are the sets of mask and non-mask labels.

Enhanced HED network: We start with ImageNet pre-trained VGG16 network [Simonyan and Zisserman, 2014] for HED, and use an improved version from [Hou *et al.*, 2017] for shadow detection. Specifically, we add two additional convolutional layers at each side output of HED. Moreover, we use larger kernel size in higher layers as suggested by [Hou *et al.*, 2017]. For the five blocks of HED, we set the kernel sizes of the newly added convolutional filters as 3x3, 3x3, 5x5, 5x5, and 7x7. We name this new version as Enhanced-HED and use it as the backbone for our proposed method.

2.2 Densely Cascaded Network via Deeply Supervised Parallel Fusion

The linear fusion of multi-scale outputs in HED does not fully capture the rich interactions between global context and local appearance. Thus, we propose a novel model that consists of two key components: (1) a deeply supervised parallel fusion network; and (2) a densely cascaded learning scheme.

Deeply Supervised Parallel Fusion network: We propose a Deeply Supervised Parallel Fusion (DSPF) network to replace the linear fusion in HED and better capture global and

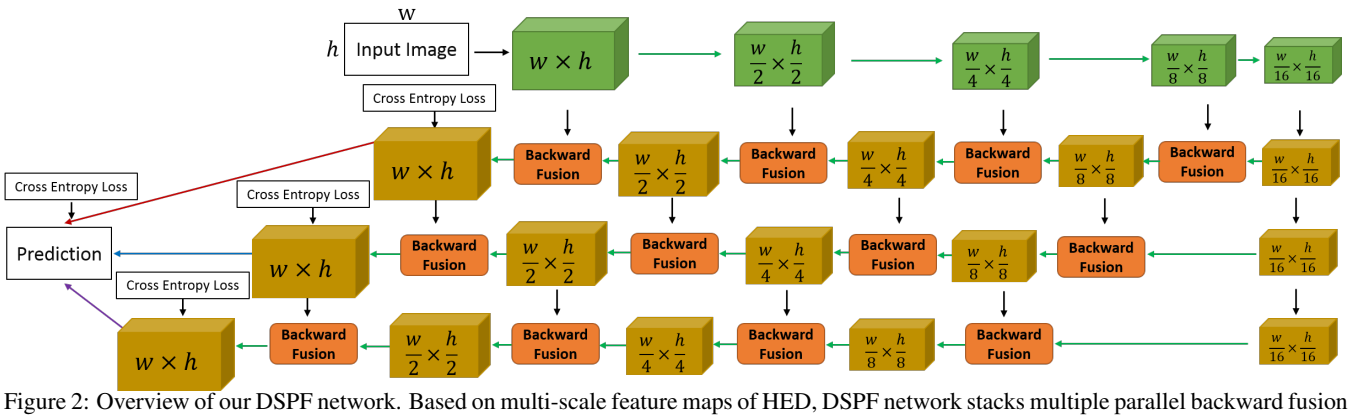

Figure 2: Overview of our DSPF network. Based on multi-scale feature maps of HED, DSPF network stacks multiple parallel backward fusion branches. DSPF network thus outputs multiple intermediate predictions via these parallel fusion pathways, with each of them supervised by its loss function. Accurate shadow segmentation is achieved by fusing these intermediate predictions (see links in red, blue, purple).

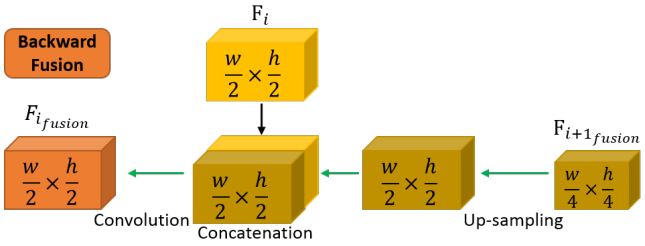

Figure 3: A basic unit of a backward fusion module in our DSPF network. This unit upsamples feature map from an upper layer and concatenate it with the feature map from current layer.

local cues. Fig 2 presents an overview of our proposed DSPF network. In order to combine global semantics with local appearance, DSPF network stacks multiple parallel fusion units, and learns to produce multi-level intermediate predictions in a deeply supervised manner. Then the final fusion of these intermediate predictions thus effectively combines both global semantic cues and local spatial details of input images.

More concretely, our DSPF network contains three parallel backward fusion branches. We observe that deeper layers in HED encode rich global semantic cues, while lower layers capture more local spatial details. To obtain a comprehensive fusion of these hierarchical feature maps, and inspired from [Pinheiro *et al.*, 2016; Wang *et al.*, 2017b], we first utilize a backward fusion branch $Branch^{(1)}$ to progressively combine these multi-scale features in a top-down manner. This fusion pathway starts from the deepest block of HED, and generates the first fusion map by transferring feature map of this deepest HED block with a 1×1 convolutional layer. The produced fusion map is then combined with the feature map in adjacent HED block via a backward fusion module. In this way, this branch progressively combines feature maps in every two adjacent blocks of HED from deeper-level layers to lower-level layers with the backward fusion module.

Specifically, we denote the fusion map from ith block as F_i. Thus, F_{i+1} is the map from upper block with reduced resolution. Our backward fusion module is given by,

$$F_{i_{fusion}} = Conv(Concat(F_i, Up(F_{i+1_{fusion}}))), \quad (3)$$

where $Up(\cdot)$ denotes the up-sampling of a feature map. $Concat(\cdot)$ means the concatenation of two feature maps which have the same resolution. $Conv(\cdot)$ refers to the 1×1 convolution operation. We also show the details of this backward fusion module in Fig 3. For these two adjacent feature maps, we first up-sample F_{i+1} to match the resolution of F_i, then we merge these two feature maps by concatenation. Finally, another convolutional layer is employed to generate final fused feature map $F_{i_{fusion}}$ of current ith block. $Branch^{(1)}$ runs from deeper layers to lower layers until the first block of HED by utilizing this backward fusion module recursively. Thus feature maps in higher-level layers are progressively combined with features in lower-level layers via these recursive fusion operations. Simultaneously, the feature resolution is also enlarged gradually. Global semantics in deeper layers are thus gradually integrated with spatial details in lower layers. Hence, the final fusion map of $Branch^{(1)}$ achieves a good fusion of the multi-scale features across the hierarchy of the base HED network. We can predict the initial response map $\hat{Y}^{(1)}$ utilizing this final fused map which has the same resolution as the input image.

$Branch^{(1)}$ computes between every two adjacent blocks at multiple stages in a top-down manner. These multi-stage fusion maps also preserve multi-scale features, which are beneficial to obtain a better fusion of global semantics and local details in input images. We thus propose to stack another backward fusion branch $Branch^{(2)}$ to further fuse these multi-stage fusion maps generated by $Branch^{(1)}$. $Branch^{(2)}$ also runs in a top-down way and combines features of every two adjacent blocks of $Branch^{(1)}$ via the backward fusion module until the last block of $Branch^{(1)}$. Similar with $Branch^{(1)}$, $Branch^{(2)}$ produces the second response map $\hat{Y}^{(2)}$ utilizing the final fusion map of this branch. Furthermore, $Branch^{(2)}$ also generates multi-scale fusion maps at multiple stages. So we stack the third backward fusion branch $Branch^{(3)}$ for further fusion. Then we can obtain the third prediction $\hat{Y}^{(3)}$ using the final fusion map.

These three intermediate shadow predictions $\hat{Y}^{(1)}$, $\hat{Y}^{(2)}$, $\hat{Y}^{(3)}$ generated by the corresponding branches $Branch^{(1)}$, $Branch^{(2)}$, $Branch^{(3)}$ receive direct supervision from multiple loss functions. With the help of this deep supervision scheme, these three intermediate predictions are enforced to capture multi-level cues. Then we can obtain the final shadow map by a weighted fusion of these these intermediate respons-

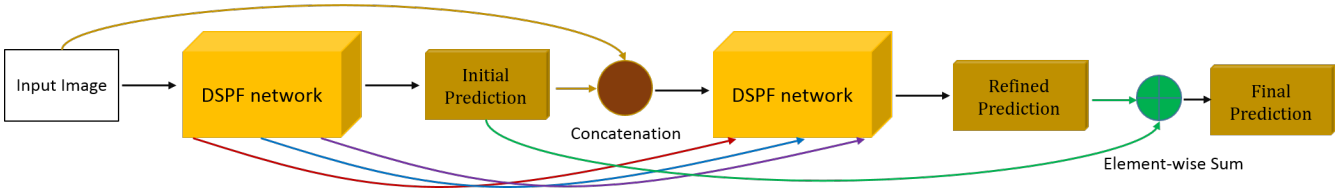

Figure 4: Illustration of the densely cascaded learning scheme. The initial prediction of the first DSPF network is concatenated with the original input image as the input of the second DSPF network. Moreover, the three intermediate responses (the three links in red, blue, purple) of the first DSPF network are summed with the three corresponding predictions in the second DSPF network. In this case, the second DSPF network learns refined the predictions by fusing these three summed intermediate response maps. Finally, the prediction of the second DSPF network is further summed with the initial prediction from the first DSPF network as the final output for shadow detection.

es:

$$\hat{Y} = w^{(1)}\hat{Y}^{(1)} + w^{(2)}\hat{Y}^{(2)} + w^{(3)}\hat{Y}^{(3)}, \quad (4)$$

where the "+" means the element-wise sum. The fusion weights $w^{(1)}, w^{(2)}, w^{(3)}$ are learned in the training phase. The loss function of DSPF network is given by

$$L_{DSPF} = L_{Branch^{(1)}} + L_{Branch^{(2)}} + L_{Branch^{(3)}} + L_{fuse}, \quad (5)$$

where the $L_{fuse} = l(Y, \hat{Y})$ is the loss between ground truth Y and prediction $\hat{Y}$, the $L_{Branch^{(1)}}$, $L_{Branch^{(2)}}$, $L_{Branch^{(3)}}$ are the losses for $Branch^{(1)}$, $Branch^{(2)}$, $Branch^{(3)}$ respectively. Note that since there are more negative non-shadow pixels than positive shadow pixels, as in HED [Xie and Tu, 2015], we use the weighted cross-entropy loss for all loss functions. In this deeply supervised scheme, the final shadow map $\hat{Y}$ learns to capture rich global semantic cues and local spatial details in an input image.

Densely cascaded learning scheme: As shown in Fig 1, global contextual cues are important to segment shadows from noisy background in natural images. In order to make the model robust to various backgrounds, we propose to integrate rich contextual cues of input images into our model. Inspired by auto-context [Tu and Bai, 2010; Li *et al.*, 2016], we propose a densely cascaded learning architecture to stack multiple DSPF networks together. This densely cascaded learning scheme specifically densely connects multiple intermediate predictions of a previous DSPF network to the next DSPF network. For a trade-off between efficiency and accuracy, we utilize two DSPF networks to form the final model via this densely cascaded learning scheme. The whole network is denoted as DC-DSPF network, and we present the details in Fig 4.

Given an input image I, the first DSPF network produces initial prediction map $\hat{Y}^{'}$, which is a weighted fusion of the three intermediate predictions $\hat{Y}^{(1)'}$, $\hat{Y}^{(2)'}$, $\hat{Y}^{(3)'}$. The second DSPF network is staked after the first one to introduce contextual cues into our model. Specifically, $\hat{Y}^{'}$ is concatenated with original input image I as the input of this second DSPF network. Moreover, these three intermediate predictions $\hat{Y}^{(1)'}$, $\hat{Y}^{(2)'}$, $\hat{Y}^{(3)'}$ of the first DSPF network are summed with three corresponding predictions $\hat{Y}^{(1)''}$, $\hat{Y}^{(2)''}$, $\hat{Y}^{(3)''}$ in this second DSPF network respectively.

$$\bar{Y}^{(m)} = \hat{Y}^{(m)'} + \hat{Y}^{(m)''}, m = 1, 2, 3. \quad (6)$$

Moreover, the second DSPF network can obtain refined prediction ($\tilde{Y}$) by

$$\bar{Y} = \bar{w}^{(1)}\bar{Y}^{(1)} + \bar{w}^{(2)}\bar{Y}^{(2)} + \bar{w}^{(3)}\bar{Y}^{(3)}. \quad (7)$$

Finally, $\bar{Y}$ is further summed with the initial estimation $\hat{Y}^{'}$ as final output of our DC-DSPF network:

$$\hat{Y} = \bar{Y} + \hat{Y}^{'}. \quad (8)$$

Through this densely connected cascade learning scheme, multi-scale feature maps encoded in the first DSPF network are effectively propagated to the second DSPF network. The whole network thus implicitly learns the spatial contextual cues in input images.

2.3 Implementation Details

We now describe our training details for DC-DSPF. All our models are trained using Caffe [Jia *et al.*, 2014] as backend. We first train HED for segmenting shadow, using the public implementation of HED [Xie and Tu, 2015]. The hyper-parameters, including the initial learning rate, weight decay and momentum, are set to $1e$-8, $2e$-4 and 0.9, respectively. Our DSPF network is initialized from the trained HED network. And our DC-DSPF is further trained on top of DSPF. The hyper-parameters of DC-DSPF are set to $1e$-8, $2e$-4 and 0.99 respectively for the initial learning rate, weight decay and momentum. All new convolutional layers are initialized with Gaussian random distribution with fixed mean (0.0) and variance(0.01). We apply random flipping for data augmentation during training.

3 Experiments and Results

We introduce our benchmark on shadow detection, and present two main results: (1) an ablation study of our method; and (2) a comparison of our method to the state-of-the-art methods. In addition, we demonstrate preliminary results of using the detected shadow regions for shadow removal.

3.1 Datasets and Evaluation Metrics

Datasets: We evaluate our method on two widely used benchmarks: SBU [Vicente *et al.*, 2016] and UCF [Zhu *et al.*, 2011]. UCF dataset consists of 245 images. Moreover, SBU dataset contains 4089 training images and 639 testing images. SBU is currently the largest and most challenging dataset for shadow detection. It covers various scenes including urban, beach, mountain, roads, parks, snow, animals, vehicles, and houses, and also contains various picture types including aerial, landscape, and close range.

Protocol: Following the evaluation protocol of [Vicente *et al.*, 2016], we train our models on SBU training set, and evaluate the trained models on SBU testing set and UCF testing

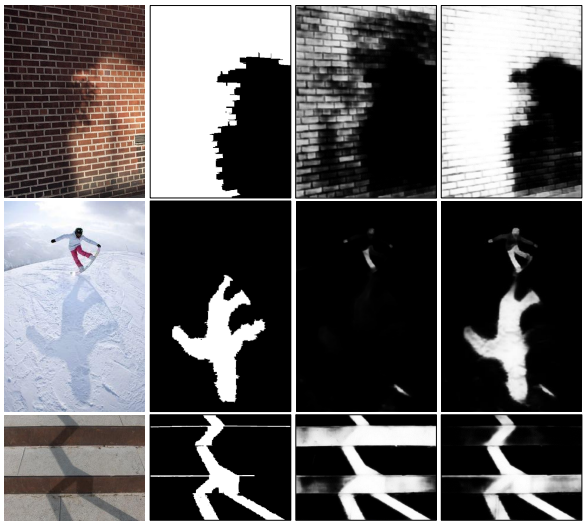

Figure 5: Visual comparison between shadow detection results of DC-DSPF and Triple-Fusion-DSPF. With the densely cascaded learning architecture, DC-DSPF captures rich contextual cues and accurately corrects the easily misclassified regions.

set. Note that testing on UCF dataset is a more challenging set as it requires a method to generalize across datasets.

Metric: We employ the Balanced Error Rate (BER) as the main evaluation metric, given by

$$BER = 1 - \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right), \qquad (9)$$

where TP,FN,TN,and FP are true positives, false negatives, true negatives, and false positives, respectively. BER is widely accepted due to the unbalanced nature of shadow data: there are much fewer shadow pixels than non-shadow ones in natural images. We also report separate per-pixel Error Rates (ER) for shadow and non-shadow pixels.

3.2 Ablation Study

We first evaluate different components of our method to better understand the proposed model. All results for our ablation study is trained on SBU training set and reported on SBU testing set. Specifically, we compare the following methods and their results are summarized in Table 1.

HED and Enhanced-HED: HED and Enhanced-HED are the backbone for our network, thus we report their results as the baseline. Compared to HED, Enhanced-HED significantly reduces the error of BER by 0.9. This result confirms the benefit of larger receptive field as suggested in [Peng *et al.*, 2017] and [Hou *et al.*, 2017].

DSPF Network: Based on Enhanced-HED, we further add our Deeply Supervised Parallel Fusion architecture (DSPF). We also vary the number of backward fusion branches from 1 to 3, denoted as Single/Double/Triple Fusion DSPF. Single fusion reduces the error of Enhanced-HED by 0.3 BER. Double and triple further reduces the error of Single-Fusion-DSPF by another 0.2 and 0.5, respectively. These results demonstrate that (1) the proposed backward fusion branch is highly effective; and (2) stacking multiple fusion branches helps to improve the performance.

	Methods	BER
HED / Enhanced-HED	HED	6.9
	Enhanced-HED	6.0
DSPF Network	Single-Fusion-DSPF	5.7
	Double-Fusion-DSPF	5.5
	Triple-Fusion-DSPF	5.2
Deep Supervision	Triple-Fusion-PF	5.6
	Triple-Fusion-DSPF	5.2
DC-DSPF	DC-Triple-Fusion-DSPF	**4.9**

Table 1: Ablation study on SBU testing set with different network architectures. The combination of the proposed parallel fusion (D-SPF) and the cascaded learning produces the best result. Our full model reduces the BER by 2.0 compared to the strong baseline of HED.

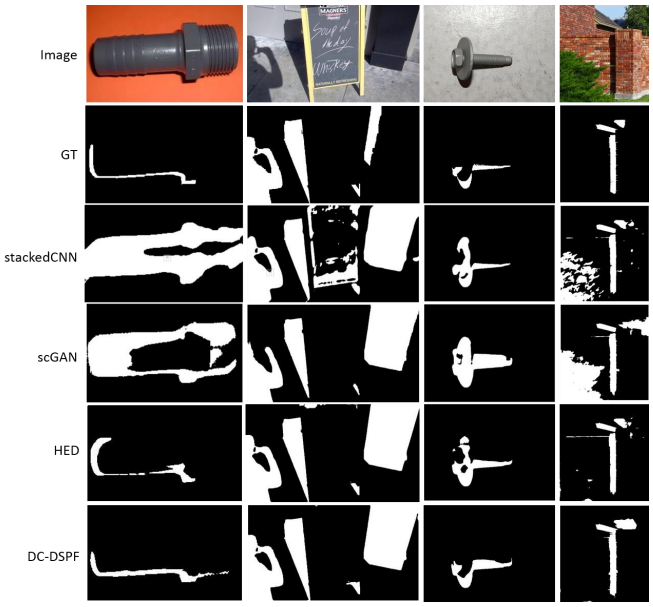

Figure 6: Visual comparison between DC-DSPF , the baseline HED and state-of-the-art stackedCNN, scGAN. DC-DSPF significantly improves the segmentation quality, captures more fine details of shadow regions and suppresses most noises from background.

Deep Supervision: We also test the effectiveness of deep supervision. Specifically, we take the Triple-Fusion-DSPF network, remove its losses for the intermediate prediction outputs at the first two backward fusion branches, and supervise the whole network only with the loss of the third backward branch. This network, called Triple-Fusion-PF (without supervision), is significantly worse than Triple-Fusion-DSPF (0.4 drop in BER). The result shows the superiority of the deeply supervised learning [Lee *et al.*, 2015].

DC-DSPF Network: Finally, we combine our best model (Triple-Fusion-DSPF) with densely cascaded learning scheme. This DC-Triple-Fusion-DSPF network further reduces the error rate by 0.3, reaching a BER of 4.9. To further understand this gap, we visualize the results of both methods in Fig 5. This result supports our design of cascaded learning.

3.3 Comparison to State-of-the-art Methods

We further compare our model of DC-DSPF network with three state-of-the-art shadow detection methods: StackedCN-N [Vicente *et al.*, 2016], scGAN [Nguyen *et al.*, 2017], and

Methods	SBU			UCF		
	BER	Per-Pixel ER		BER	Per-Pixel ER	
		Shad.	Non Shad.		Shad.	Non Shad.
StackedCNN	11.0	9.6	12.5	13.0	9.0	17.1
scGAN	9.1	7.8	10.4	11.5	7.7	15.3
DSC	5.6	-	-	8.1	-	-
HED	6.9	7.0	6.9	9.5	7.7	11.4
DC-DSPF	**4.9**	**4.7**	**5.1**	**7.9**	**6.5**	**9.3**

Table 2: Comparison to state-of-the-art methods on SBU and UCF testing sets. DC-DSPF consistently outperforms previous methods.

Methods	SBU	UCF
PSPNet	8.6	11.8
Amulet	15.1	15.2
SRM	7.3	9.8
HED	6.9	9.5
Our DSPF	**4.9**	**7.9**

Table 3: Comparison to state-of-the-art dense labeling methods for semantic segmentation (PSPNet) and saliency detection (Amulet, SRM). Results are reported on SBU and UCF testing sets. Even simple HED already outperforms PSPNet, Amulet, and SRM. Our DC-DSPF further reduces the BER.

DSC [Hu *et al.*, 2018] on SBU and UCF testing set. Moreover, we include the HED network as a baseline. The results are presented in Table 2.

Surprisingly, the baseline of HED already outperforms stack-CNN and scGAN. Our DC-DSPF further improves the performance of HED and outperforms the state-of-the-art methods. For BER, DC-DSPF obtains a significant error reduction by 46% and 31% than scGAN on SBU and UCF dataset, respectively. Moreover, our method can consistently reduce the error in shadow pixels (40% on SBU and 16% on UCF) and detect more none shadow pixels (51% on SBU and 39% on UCF) in comparison to scGAN. Our DC-DSPF also obtains a clear error reduction than previous best performing DSC by 0.7 BER on SBU and 0.2 BER on UCF.

We also present visual comparison between the results of our method and the state-of-the-art methods as shown in Fig 6. DC-DSPF preserves rich contextual cues and achieves more precise estimation on the easily misclassified regions where global cues should be taken into consideration.

Moreover, similar to [Hu *et al.*, 2018], we also compare our DC-DSPF with methods for semantic segmentation (PSPNet [Zhao *et al.*, 2017]) and saliency detection (Amulet [Zhang *et al.*, 2017], SRM [Wang *et al.*, 2017a]) on the task of shadow detection in Table 3. Even the baseline HED significantly surpasses PSPNet, Amulet, and SRM on SBU and UCF. Our DC-DSPF further increases the performance gap. The results demonstrate the superior performance of our method. Furthermore, these results also suggest the fundamental difference in the modeling of semantic segmentation, saliency detection and shadow detection. These three tasks are all formulated as dense labeling of pixels. However, the modeling of these tasks differs significantly. Semantic segmentation requires the reasoning of global context between objects and scene components [Zhao *et al.*, 2017]. This is not required for saliency detection and shadow detection. In comparison to saliency detection [Zhang *et al.*, 2017; Wang *et al.*, 2017a], shadow detection focuses on the local properties of image regions (e.g., colors and textures) and does not require explicit modeling of objects.

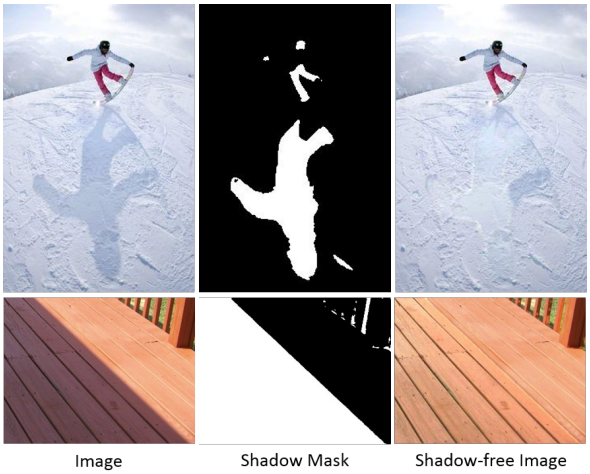

Figure 7: Qualitative examples of single image shadow removal using shadow masks produced by our method.

3.4 Results of Shadow Removal

Finally, we make use of our output shadow mask for the application of shadow removal in a static image. We employ the method in [Corina B. *et al.*, 2011]–a simplified version of [Guo *et al.*, 2011]. As [Corina B. *et al.*, 2011] denotes that shadow detection remains a challenge for obtaining shadow-free images. The performance of shadow detection results directly influences the quality of shadow removal results. We supply the shadow mask generated by our DC-DSPF network and original image as the input of this method, and produce the shadow-free results. Qualitative examples are shown in Fig 7. These examples suggest that our precise shadow detection enables good shadow removal results.

4 Conclusion

In this work, we consider the challenging problem of shadow detection. We propose a novel DC-DSPF network that combines a DSPF network and a densely cascaded learning architecture. The DSPF network stacks multiple parallel fusion branches, and learns a comprehensive fusion of global semantic cues and local spatial details in a deeply supervised way. Moreover, the densely cascaded learning scheme helps to capture rich contextual cues. Experimental results show that our DC-DSPF network significantly outperforms state-of-the-art methods on major benchmarks.

Acknowledgements

This work is partial funded by the National Key Research and Development Program of China (Grant No.2016YFB1001005), the National Natural Science Foundation of China (Grant No.61673375, Grant No.61602485), and the Projects of Chinese Academy of Science (Grant No.QYZDB-SSW-JSC006, Grant No.173211KYSB20160008).

References

[Corina B. *et al.*, 2011] Peter J. K. Corina B., Zoltan B., and Laszlo V. Shadow detection and removal from a single image. *SSIP*, 2011.

[Ecins *et al.*, 2014] Aleksandrs Ecins, Cornelia Fermuller, and Yiannis Aloimonos. Shadow free segmentation in still images using local density measure. In *ICCP*, 2014.

[Finlayson *et al.*, 2006] Graham D. Finlayson, Steven D. Hordley, Cheng Lu, and Mark S. Drew. On the removal of shadows from images. *TPAMI*, 28(1), 2006.

[Finlayson *et al.*, 2009] Graham D. Finlayson, Mark S. Drew, and Cheng Lu. Entropy minimization for shadow removal. *IJCV*, 85(1), 2009.

[Guo *et al.*, 2011] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Single-image shadow detection and removal using paired regions. In *CVPR*, 2011.

[Hou *et al.*, 2017] Qibin Hou, Ming Ming Cheng, Xiao Wei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. In *CVPR*, 2017.

[Hu *et al.*, 2018] Xiaowei Hu, Lei Zhu, Chi Wing Fu, Jing Qin, and Pheng Ann Heng. Direction-aware spatial context features for shadow detection. In *CVPR*, 2018.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[Karsch *et al.*, 2011] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. In *SIGGRAPH Asia Conference*, 2011.

[Khan *et al.*, 2016] Salman H. Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Automatic shadow detection and removal from a single image. *TPAMI*, 2016.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[Lalonde *et al.*, 2010] Jean Francois Lalonde, Alexei A. Efros, and Srinivasa G. Narasimhan. Estimating natural illumination from a single outdoor image. In *ICCV*, 2010.

[Lee *et al.*, 2015] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *AISTATS*, volume 2, page 6, 2015.

[Li *et al.*, 2016] Ke Li, Bharath Hariharan, and Jitendra Malik. Iterative instance segmentation. In *CVPR*, 2016.

[Nguyen *et al.*, 2017] Vu Nguyen, Tomas F Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *ICCV*, 2017.

[Panagopoulos *et al.*, 2012] Alexandros Panagopoulos, Chaohui Wang, Dimitris Samaras, and Nikos Paragios. Simultaneous cast shadows, illumination and geometry inference using hypergraphs. *TPAMI*, 35(2), 2012.

[Peng *et al.*, 2017] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters – improve semantic segmentation by global convolutional network. 2017.

[Pinheiro *et al.*, 2016] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *ECCV*, 2016.

[Qu *et al.*, 2017] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W. H. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *CVPR*, 2017.

[Shelhamer *et al.*, 2017] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 2017.

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

[Tu and Bai, 2010] Zhuowen Tu and Xiang Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *TPAMI*, 2010.

[Tumblin and Williams, 2011] Jack Tumblin and Lance Williams. What characterizes a shadow boundary under the sun and sky? In *ICCV*, 2011.

[Vicente *et al.*, 2016] 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*. Springer, 2016.

[Vicente *et al.*, 2018] Tomas F Yago Vicente, Minh Hoai, and Dimitris Samaras. Leave-one-out kernel optimization for shadow detection and removal. *TPAMI*, 2018.

[Wang *et al.*, 2017a] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, 2017.

[Wang *et al.*, 2017b] Yupei Wang, Xin Zhao, and Kaiqi Huang. Deep crisp boundaries. In *CVPR*, 2017.

[Xie and Tu, 2015] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015.

[Zhang *et al.*, 2017] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Ruan Xiang. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, 2017.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[Zhu *et al.*, 2011] Jiejie Zhu, Kegan G. G. Samuel, Syed Z. Masood, and Marshall F. Tappen. Learning to recognize shadows in monochromatic natural images. In *CVPR*, 2011.