Defocus Blur Detection via Depth Distillation

Xiaodong Cun and Chi-Man Pun

University of Macau, Macau, China
{yb87432, cmpun}@umac.mo

Abstract. Defocus Blur Detection (DBD) aims to separate in-focus and out-of-focus regions from a single image pixel-wisely. This task has been paid much attention since bokeh effects are widely used in digital cameras and smartphone photography. However, identifying obscure homogeneous regions and borderline transitions in partially defocus images is still challenging. To solve these problems, we introduce depth information into DBD for the first time. When the camera parameters are fixed, we argue that the accuracy of DBD is highly related to scene depth. Hence, we consider the depth information as the approximate soft label of DBD and propose a joint learning framework inspired by knowledge distillation. In detail, we learn the defocus blur from ground truth and the depth distilled from a well-trained depth estimation network at the same time. Thus, the sharp region will provide a strong prior for depth estimation while the blur detection also gains benefits from the distilled depth. Besides, we propose a novel decoder in the fully convolutional network (FCN) as our network structure. In each level of the decoder, we design the Selective Reception Field Block (SRFB) for merging multi-scale features efficiently and reuse the side outputs as Supervision-guided Attention Block (SAB). Unlike previous methods, the proposed decoder builds reception field pyramids and emphasizes salient regions simply and efficiently. Experiments show that our approach outperforms 11 other state-of-the-art methods on two popular datasets. Our method also runs at over 30 fps on a single GPU, which is 2x faster than previous works. The code is available at: https://github.com/vinthony/depth-distillation

Keywords: Defocus Blur Detection, Attention Module, Knowledge Distillation

1 Introduction

Defocus blur, which is also called the bokeh effect in photography, has been widely used in everyday photos. The focus region emphasizes the salient object while the out-of-focus blur can protect the privacy of people appearing in the photo. Moreover, detecting this kind of blur is also crucial since the detected defocus region could be potentially useful in performing tasks. Such tasks include auto-refocus[1], salient object detection [14] and image retargeting [15].

Since DBD has a long history in computer vision [32, 27, 40, 32, 8], traditional methods focus on designing novel hand-crafted features such as the gradient [8,

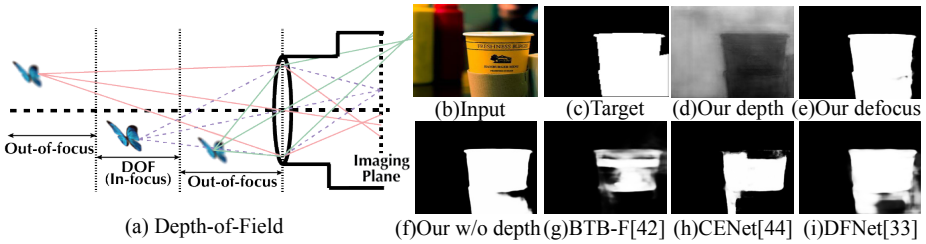


Fig. 1: We first leverage depth into DBD since predicting defocus blur is similar to estimating the *Depth-of-Field*(DOF) from a partial defocus image as in (a). By involving depth in DBD network, our method assumes that the depths in DOF regions are similar(d) and that the region with more similar depths might be part of DOF/out-of-focus region as well(e). Thus, we obtain more accurate results than other DBD methods(f)-(i).

39] or the frequency domain features [32, 33, 40]. However, these methods extract limited features and lack high-level semantic information. Thus, if the scene is complex, it is hard to discriminate the defocus region by particular features.

Recently, deep learning-based methods have shown superior performance in various computer vision tasks as well as defocus blur detection. For example, Park *et al.* [27] train a CNN to classify the sharpness of each local patch in an image. Deeper fully convolutional methods [45, 43, 44, 34] have been proposed by regarding the DBD as scene segmentation. Although these methods emphasize the importance of image scales in DBD, they are still considering DBD from a 2D perspective and rely solely on the power of the datasets and neural network.

In this paper, we start from the cause of defocus blur in the photography. As shown in Fig. 1(a), the sharp focus region, also called the depth-of-field(DOF[35]), is formed because the camera only images clear photo in a certain depth range[1]. When the light waves intersect behind or in front of the imaging plane (red and green lines in Fig. 1(a)), the area they originated from will be blurred in final image. Since the homogeneous region in DBD often includes multiple objects and since it is difficult to be detected by edges or semantic features, the distance between the camera and scene objects (depth) provides a strong prior for classification. However, the unconstrained depth estimation is an ill-posed problem. To evaluate on currently available DBD datasets and provide fair comparison with previous methods, we propose depth distillation by using a pre-trained network [3] as regularization and learn the defocus map simultaneously. In addition, we design a Supervision-guided Attention Block (SAB) for re-weighting the learned features based on each level of side outputs. Finally, the blur confidence is relative, which means *a sharp patch can be regarded as blurry when we enlarge it and vice versa.* Although previous methods [43, 44, 34] have discussed it by multi-stream or cross-layer fusion networks, we consider it in an

[1] We simplify this model by ignoring the influence of camera parameters since we can only obtain a 2D RGB image in the dataset.

efficient way by designing Selective Reception Field Block (SRFB) in each decoder. Our block extracts larger reception fields to build richer feature pyramids and uses a global selective attention to weight the importance of useful features. By involving depth estimation into DBD and the proposed blocks, our network outperforms other methods on the defocus detection. As shown in Fig. 1 (b)-(i), previous methods for DBD are sensitive to color, while in our network, DBD and depth estimation tasks build on each other and predict the results perfectly.

We summarize the contributions of this paper as follows:

- To the best of our knowledge, this is the first attempt to introduce depth information in DBD and distill the knowledge of pre-trained depth model as regularization of DBD network.
- In each decoder of our framework, we design the Supervision-guided Attention Block (SAB), which reuses the side depth and defocus map for spatial attention. Considering the sensitivity of scale, we also design the Selective Reception Fields Block (SRFB) to extract multiple reception field features.
- We conduct the experiments on two popular DBD benchmarks with 11 state-of-the-art methods (7 from DBD and 4 from related tasks). The results show that our proposed method can achieve much better results.

2 Related Works

Traditional methods Out-of-focus and DOF regions have significant visible differences in sharpness. Thus, traditional DBD methods are designed based on identifiable hand-crafted features such as gradient or edge representation [16, 32]. For example, Yi *et al.* [40] use local binary patterns as focus sharpness metric. Shi *et al.* [33] use sparse representation to correlate the sparse edges and blur strength. Frequency-based methods are another noticeable trend in hand-crafted features, since the high-frequency components of the in-focus region and out-of-focus region are different. For example, Golestaneh *et al.* [8] use multi-scale high-frequency fusion and sort transform to determine the magnitudes of gradients. Although the methods based on hand-crafted features have been demonstrated to be effective in some cases, these methods are not robust enough in a broader variety of complex scenes.

Learning-based methods Deep neural networks, especially CNNs, are widely used in many computer vision and image processing tasks. Park *et al.* [27] propose the first CNN based method to DBD by combining the hand-crafted features and pre-trained deep features together. In this method, the image is scanned in a patch-by-path manner to find the defocus blur. Inspired by the object detection and segmentation methods, Zhao *et al.* [43, 44] firstly use the full convolutional network-based method by considering DBD to be sensitive to scale. Following this idea, Tang *et al.* [34] design a novel network structure for feature fusion and Zhao *et al.* [45] ensemble multiple networks to enhance diversity. In contrast to previous studies, Lee *et al.* [18] address the lack of datasets by learning from synthesized rendered dataset with domain adaptation. However, previous

learning-based methods only focus on learning with stronger networks[45, 44, 43] or dataset [18].

Depth estimation and depth-assisted methods Estimating the depth from a single image is ill-defined since inferring 3D information requires multi-views. However, monocular depth estimation in restricted scenes is possible, for example, with indoor scenes [5] or the road in a driving context [6, 7]. In contrast, predicting the depth in the wild is still a challenge. Chen *et al.* [3] propose an end-to-end network based on point relationships. However, this network only predicts relations between the objects other than absolute depth. Li *et al.* [20] generate the multi-view disparity of humans from video of people who are frozen in place, and this task only works when the person are in the scene. Depth also plays an important role in other tasks. Most methods consider the depth to be known by the sensor. Such as RGB-D object detection [29] and RGB-D salient object detection [28]. Some methods exploit the knowledge of depth in related tasks, such as depth-assisted view synthesis [4] and depth attentional features for deraining [12].

3 Methods

We define DBD as a supervised pixel-wise binary classification problem. Rather than considering the defocus region as positive, we learn the opposite DOF (in-focus) region as previously [43–45]. Given the input image I and the corresponding ground truth DOF region M, we construct a deep convolutional network $\Phi(\cdot;\theta)$ by feeding the image I to generate the DBD maps $\Phi_{df}(I)$ and depth maps $\Phi_{dp}(I)$. Then, we optimize the parameters θ of Φ to minimize the the defocus metrics L_{df} and depth metrics L_{dp}:

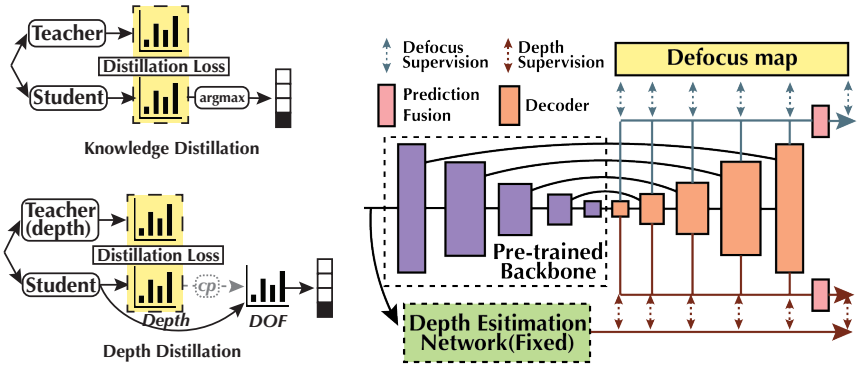
$$\arg\min_{\theta} L_{df}(M, \Phi_{df}(I;\theta)) + L_{dp}(\Re(I), \Phi_{dp}(I;\theta)) \tag{1}$$

where $\Re$ is a pre-trained depth estimation network[3] for depth distillation. Below, we introduce the details of depth distillation, network structure and metrics.

3.1 Depth Distillation

In general, knowledge distillation [9, 24] aims to transfer the knowledge for network structure optimization. In detail, as shown in Fig. 2(a), they regularize the compact (student) model using a larger (teacher) network in the space of **continuous** soft label(the output of Softmax), other than transferring the knowledge using predicted **discrete** hard targets.

Interestingly, we find that DBD(**discrete, classification task**) and depth estimation (**continuous, regression task**) have a similar relationship with that between hard and soft labels in knowledge distillation. In photography, the sharp focus region (DOF) is mathematically defined as [25]: $DOF \approx \frac{2NCD^2}{f^2}$, where N is the F-number of lenses, C is the circle of confusion and f is the focal length, respectively. The depth D is the only one which is not the camera



(a) Depth Distillation (b) Network Structure

Fig. 2: (a)Comparison with knowledge distillation and the proposed depth distillation. (b)Our network structure. Under FCN framework, we distill the depth information from a pre-trained depth estimation network[3] and design novel decoders for DBD and depth estimation jointly.

parameter (*cp*). Thus, as shown in Fig. 2(a), we propose *depth distillation* to help defocus blur detection. In detail, we consider that the depth is the approximate soft label and distill the depth information from a pre-trained network as regularization of DBD. Instead of calculating the DOF from depth map directly and inferring the defocus map as knowledge distillation, our network can predict the defocus map and distill depth jointly because the camera parameters are unavailable. Although the structure of depth distillation and knowledge distillation are similar, the goal is totally different: We aim to involve the 3D information into DBD task other than distilling a compact model from teacher network. For implementation, we design a simple yet effective framework to achieve previous analysis. As shown in Fig. 2(b), we generate multiple outputs for depth estimation, which are supervised by a pre-trained network. Then, we fuse all the side outputs to obtain the final depth through a fusion (1x1 Conv.) block. However, single image depth estimation is ill-posed since the dense depth is hard to be collected especially in unconstrained settings. Thus, we choose the relative depth network (Chen *et al.* [3]) as teacher network. Specifically, they aim to learn the relationships between scene objects other than accurate depth values. Thus, they label the spatial relationship between 800 pairs of points (e.g., point A, B share the same depth, A is closer to camera than B and vice versa) pre-image as the supervision. Then, the neural network can predict the dense relative depth with the help of large-scale training samples.

Leveraging depth information to DBD as depth distillation has many benefits. First, the depth distillation helps our network to understand the scenes better except for the binary classification (similar to the relationships in knowledge distillation as discussed). Then, the blurriness region in the input also gives

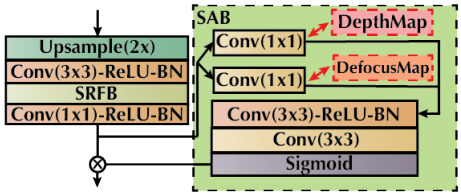


Fig. 3: The detailed structure of the proposed decoder, where the red arrows mean defocus supervision and depth distillation, respectively.

a dense hint to relative depth estimation. Finally, by depth distillation, we do not need the pre-trained depth network in testing, which also helps to build an efficient algorithm. Distilling from relative depth network is also critical. Since the training dataset of DBD only contains 600 images, the pre-trained relative depth network(421K training images in the wild) involves more accurate 3D features from larger-scale datasets to our network and task. Besides, we find that the network of Chen *et al.* automatically locates the salient object and predicts its relative depth. Luckly, DBD has a similar goal because the photographers often use the defocus blur to stress the important views.

3.2 Network Structure

Our network structure is based on *Fully Convolutional Networks* (FCN [23]). As shown in Fig. 2(b), we extract multi-scale features (5 layers in total) before each MaxPooling layer in a pre-trained ResNeXt101 [38] on ImageNet. These multi-scale features contain both high-level semantic features and low-level details for further detection. In each decoder, as shown in Fig. 3, we use the upsampling layer with convolution instead of deconvolution layer (or transpose convolution layer) to avoid checkerboard artifacts [18, 26]. Then, by considering the importance of scale in DBD, we proceed using several aspects of multi-scale feature modeling and preservation. On the one hand, we design auxiliary classifiers in each level of the decoder as in [10, 21, 18] to prevent over-fitting and to generate multi-scale results. Differently, in each level of the decoder, we design two auxiliary classifiers for the supervision from DBD and depth distillation, respectively. Each auxiliary classifier is defined as a 1x1 convolution layer for side prediction, and we reuse these side outputs as the Supervision-guided Attention Block (SAB) for spatial attention (as shown in Fig. 3). Then, the final defocus and depth map can be generated by merging all multi-scale intermediate output maps with a 1x1 convolution layer as the PredictionFusion block in Fig.2(b). On the other hand, we model multi-scale reception fields in each level of the decoder and propose the Selective Reception Field Block (SRFB) for efficiently selecting and merging the features in multi-contexts. Next, we provide the details of the proposed blocks.

Supervision-guided Attention Block Inspired by recently proposed attention mechanisms [11, 36], we increase the non-linearity of network with the at-

tention block. In detail, we generate the attention map from the side outputs since it also has a stronger prior knowledge for further feature weighting. As shown in Fig. 3, after the supervision of the auxiliary classifier, we feed the auxiliary outputs of DBD and depth to the network again. Then we generate the spatial attention by two convolution blocks and a Sigmoid function. Finally, we multiply the original features by the generated attention map. These attentions rescale the features spatially before the next decoder.

Selective Reception Field Block Since DBD needs to deal with scale carefully, previous works [43, 44, 34] merge multiple networks with multi-scale inputs, or recurrently and crossly fuse multi-scale features. However, these networks are still heavy and computationally inefficient. Rather than designing multi-stream networks or fusing by cross layers, we design an efficient multi-branch block for the extraction of multiple reception fields in each individual decoder.

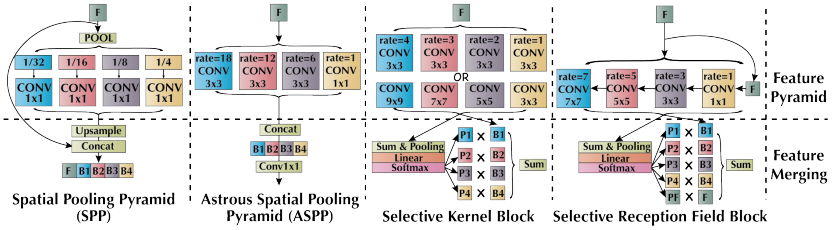


Fig. 4: Different types of multi-kernel feature pyramids. The proposed Selective Reception Field Block enlarges the reception fields of Selective Kernel Block using larger reception field pyramids. F repersents the original feature, Bx and Px are the x-th branch and the corresponding probability, respectively. rate = k means the dilation rate equals to k.

This is a natural way to extract multi-scale features using different kernel sizes. For example, the widely used (atrous) spatial pooling pyramid (SPP [42] or ASPP [2]) has been successful in semantic segmentation and other related tasks [41, 31]. More recently, *Selective Kernel Networks* (SK-Block [19]) have been proposed for weighting multiple kernels in image classification. As shown in Fig. 4, we find that the SK-Block has a similar purpose to (A)SPP. Thus, we give a general formulation of these blocks by modeling them as a two stage process: *feature pyramid* and *feature merging*. The (A)SPP extracts the multi-context features by pooling or dilated convolution and then merges with convolutional block. In contrast, the SK-Block models the multi-context features using different convolution kernels (or convolution with different dilation rates). Then, it produces the probability of each branch using the global attention and uses it to weight each kernel.

However, if we insert SK-Block to FCN directly, the reception fields are still local and enlarging it needs much more memory. Thus, inspired by the (A)SPP, we design the Selective Reception Field Block with the following improvements

to SK-Block: First, we add the original feature into feature pyramid and merging. By involving the original feature, other branches will try to learn the residuals of input. On the other hand, we aim to create richer and larger receptive fields in the feature pyramid. In detail, we use a sequence of convolutional layers with the growth of dilation and convolutional kernel together as feature pyramid, which is inspired by *Reception Field Block* (RFB)[22]. Note that RFB are proposed for object detection as an inception-like structure. In contrast, we build the feature pyramid, which is inspired by their intentions, and use these blocks as the decoders in the FCN framework. Thus, we have enlarged the reception field of the SK-Block substantially, which contains multi-scale features for weighting and selection. For example, when there are four branches in the block (as shown in Fig. 4), the reception field of the Sk-Block is 11 (9×9 Conv. or 3×3 Conv. dilation=4) while ours is 43 (7×7 Conv. with dilation=7).

3.3 Loss Function

Our training loss is defined as a combination of overall auxiliary supervisions and the final prediction of defocus estimation and depth distillation.

For defocus estimation, we use the weighed binary cross entropy (BCE) loss as in [13, 46] for all the auxiliary outputs $M'_k(k \in [1, .., 5])$ and the final output M_f compared with ground truth M: $\ell_{defocus} = \ell_{bce}(M, M_f) + \sum_k \alpha^k \ell_{bce}(M, M'_k)$, where the weighted BCE is defined as:

$$\ell_{bce}(M, M') = -(1 - \frac{TP}{N_p})Mlog(M') - (1 - \frac{TN}{N_n})(1 - M)log(1 - M') \quad (2)$$

TP and TN are the numbers of true positives and true negatives in the samples, N_p and N_n are the numbers of in-focus and out-of-focus pixels, respectively.

As for the depth distillation, giving the pre-trained depth estimation network as Φ_{rd}, the input image I and the predicted depth $\Phi_d(I)$ in our network, we define the depth distillation loss in level k as:$\ell^k_{depth} = ||\Phi^k_d(I) - \Phi_{rd}(I)||_2$. Similar to defocus estimation, our network predicts multi-scale depth output and fuses the side outputs to generate the final results. Thus, the full loss of depth distillation can be written as:

$$\ell_{depth} = \ell^f_{depth} + \sum_k \beta^k \ell^k_{depth} \quad (3)$$

where ℓ^f_{depth} represents the results after the final fusion layer and ℓ^k_{depth} represents the k levels of auxiliary outputs.

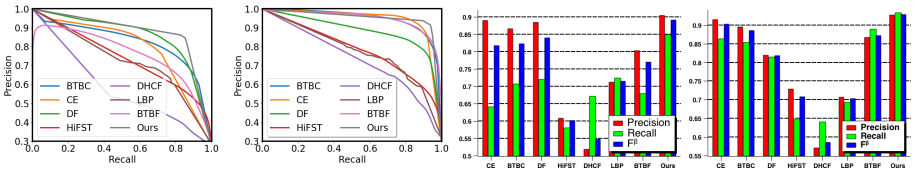
Overall, the total loss of our network is: $L = \ell_{defocus} + \gamma \ell_{depth}$. All the α, β are experimentally set to 1, and γ equals to 0.1.

4 Experiments

Implementation Details We implement our method in the PyTorch framework. The parameters of the encoder backbone are initialized from the pre-trained ResNeXt101 [38] on ImageNet, while the other parameters are random

Table 1: Comparisons with state-of-the-art methods on F^β and MAE score. We compare our method with 7 DBD methods [40, 27, 8, 43, 44, 34, 45] and 4 methods on the related tasks(salient object detection [30, 37] and shadow detection [47, 13]). Our method achieves the best performance over 11 methods on two datasets. Meanwhile, our method is 2x faster than previous DBD methods.

Datasets	Metrics	[40]	[27]	[8]	[43]	[44]	[34]	[45]	[13]	[47]	[30]	[37]	Ours
CUHK	F^β	.787	.477	.772	.867	.889	.818	.906	.898	.912	.922	.901	**.927**
100	MAE	.136	.372	.219	.107	.082	.117	.059	.057	.046	.049	.055	**.042**
DUT	F^β	.719	.468	.687	.761	.827	.823	.817	.844	.877	.827	.866	**.891**
500	MAE	.193	.410	.248	.194	.138	.118	.135	.109	.080	.120	.092	**.073**
-	FPS	.11	.09	.02	.04	.08	17.9	15.6	40.0	22.2	**90.9**	50.0	35.7



(a) DUT500 (b) CUHK100 (c) DUT500 (d) CUHK100

Fig. 5: (a) and (b) are *Precision-Recall Curves*, and (c) and (d) are the comparison of Precision, Recall and F^β on two datasets. The proposed method achieves the best performance on various metrics.

noise. We utilize the Stochastic Gradient Descent (SGD) algorithm to optimize the network with momentum of 0.9 and learning rate of 0.005. We resize all the images to 320x320 for training and evaluating the results in the same resolution as previous. Our network is trained on a computer equipped with an Intel 3.60 GHz CPU, 32G memory and a single GTX 1080 GPU. We set the batch size equals to 6, and the whole training process takes less than 2 hours. Regarding interference, our network can generate a 320x320 image in 0.028s (**35.7 fps**), which is faster than previous DBD methods as shown in Table 1. Note that, for the training, we do not use any additional samples [44] or synthesized samples [18]. Additionally, for fair comparison, all the results are raw outputs from the network without any post-processing (such as dense conditional random fields [17]). More comparisons and experiments can be found in the supplementary materials.

Dataset We evaluate our algorithm on two publicly available datasets for DBD. The first is the CUHK dataset [32], which contains 704 images with partially defocus blur. Another dataset is the DUT dataset [43], which contains 500 difficult samples with obscure homogeneous, low-contrast focal regions and background clutter. We train our network on the same split of 604 images from the CUHK dataset as previous work [43–45] and test on the remaining 100 images (CUHK100) and the whole DUT dataset(DUT500).

Fig. 6: Comparison with state-of-the-art DBD methods. From the left to right is: (a) Input, (b) Target, (c) Ours, (d) BTB-C [44], (e) BTB-F [43], (f) CENet [45], (g) LBP [40],(h) DHCF [27],(i) HiFST [8] and (j) DFNet [34]. Our method generates more convincing DBD maps than others.

Metrics We evaluate DBD on three aspects as previous works. The first metric is the *Precision-Recall (PR) Curve* for binary classification accuracy. All the results are normalized to [0, 255] and given a threshold in each integer interval. Second, we compute the mean precision, recall and F-measure scores (F^β) on the binarized results by an adaptive threshold. The threshold is determined by the 1.5 times of the mean pixel value. The F-measure is defined as:$F^\beta = \frac{(1+\beta^2)\times Precision\times Recall}{\beta^2\times Precision+Recall}$, where $\beta^2 = 0.3$ and $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{FN+TP}$, respectively. A larger F^β indicates a better result. Last, we report the mean absolute error (MAE) for the average pixel differences between the ground truth M and predicted M'. MAE is defined as: $MAE = \frac{1}{WH}\sum_{x=0}^{W}\sum_{y=0}^{H}|M(x,y)-M'(x,y)|$, where W, H are the width and height and x, y are the spatial coordinates of the image, respectively.

4.1 Comparisons with State-of-the-Art Methods

We compare our algorithm with several state-of-the-art methods, including deep learning-based methods for DBD, such as: deep and hand-crafted features based method (DHCF [27]), multi-stream bottom-top-bottom (BTB-F [43], BTB-C [44]), network cross-ensemble(CENet [45]) and the network with recurrently feature reuse and fusion (DFNet [34]). In addition, we also conduct the experiments on state-of-the-art hand-crafted feature based methods, including local binary patterns (LBP [40]) and high-frequency multi-scale fusion and sort transform of gradient magnitudes (HiFST [8]). Note that, all the predicted maps of DBD come from the author's website or the public implementation with recommended hyper-parameters for fair comparison. For there are few learning-based DBD methods, we also compare our methods with 4 state-of-the-art learning-based

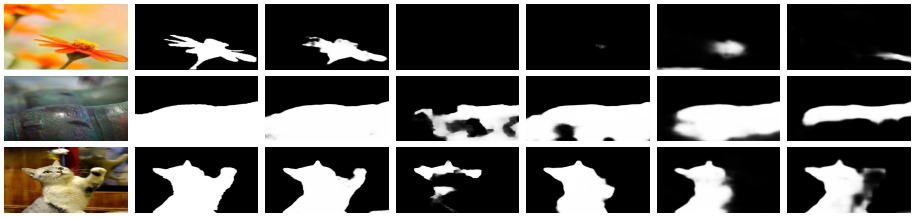


Fig. 7: The produced DBD map of our method outperforms others state-of-the-art network structures on related tasks. From the left to right is: (a)Input, (b)Target, (c)Ours, (d)BASNet[30], (e)BDRAR[47], (f)CPD[37], (g)DSC[13].

methods on some relevant tasks: such as bidirectional feature pyramid network with recurrent attention (BDRAR [47]) and direction-aware attention (DSC [13]) for shadow detection, boundary-aware loss (BAS [30]) and cascaded partial decoder (CPD [37]) for salient object detection. All the networks of relevant tasks are trained on our framework with the same input resolution and batch size.

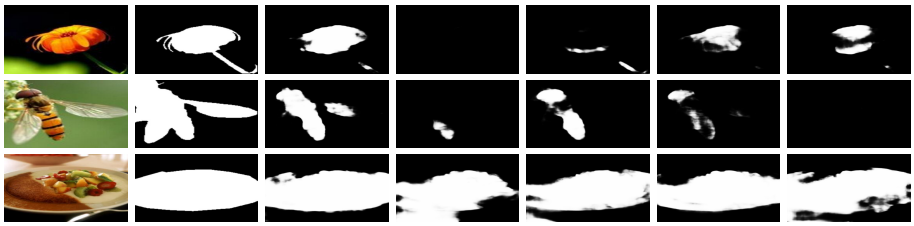
We illustrate the numerical comparison of our method and state-of-the-art methods on two public datasets in Table 1 and Fig. 5. It is clear that our method outperforms others with a larger margin on all numerical metrics. The results show that our network with depth and multi-scale features understands the complex scenes well. We also give some visual samples to compare with state-of-the-art DBD methods in Fig. 6. Our methods also show the superior visual quality. Apart from the great object awareness in examples, our network also predicts the homogeneous regions well (such as the plane in the fourth example) because of depth distillation. For comparison with related tasks, Table 1 also gives a clear result. Our network has better performance in DBD than the boundary awareness network BASNet [30] or direction awareness DSC [13] because depth is more important in our task. For example, boundary loss in BASNet [30] is benefit on CUHK100 (As Table 1) but worse in DUT500 because the homogeneous regions in DUT500 are not related to edge. As shown in Fig. 7, our method can achieve much better results than the other methods.

4.2 Ablation Studies of Network Structure

Backbone choice We choose different backbones for our network structure, especially the widely used VGG19 in previous work and ResNeXt101. For the ablation study, we use the FCN [23] with auxiliary outputs and ResNeXt101 as feature extractor and compare with our main contributions in Table 2. Since the CUHK100 dataset is small and simple, the metric differences on this dataset is not too large. While on the DUT500 dataset, ResNeXt101 can extract richer features and gain much better results. By comparing with the other state-of-the-art methods on DBD (7 DBD methods [40, 27, 8, 43, 44, 34, 45] in Table 1,

Table 2: Ablation study. The first two experiments use VGG19 as feature extractor while the last five experiments use ResNeXt101 as feature extractor. OursFull means the FCN+D+SRFB+SA.

Datasets	Metrics	FCN VGG	OursFull VGG	FCN ResNeXt	+D	+D +SRFB	+D+RFB +SA	OursFull ResNeXt
CUHK	F^β	0.911	0.919	0.917	0.922	0.926	**0.931**	0.927
100	MAE	0.053	0.048	0.046	0.046	0.045	**0.040**	0.042
DUT	F^β	0.800	0.844	0.879	0.883	0.888	0.887	**0.891**
500	MAE	0.148	0.113	0.080	0.077	0.076	0.075	**0.073**



(a) (b) (c) (d) (e) (f) (g)

Fig. 8: Ablation study of network structure. From left to right is: (a) input, (b) Target (c) defocus+D+SRFB+SA (Our Full method) (d) defocus only, (e) defocus+D, (f) defocus+D+SRFB, (g) defocus+D+RFB+SA.

and 4 related tasks in Table 1), our network can also improve the performance significantly on the similar pre-trained VGG19 backbone.

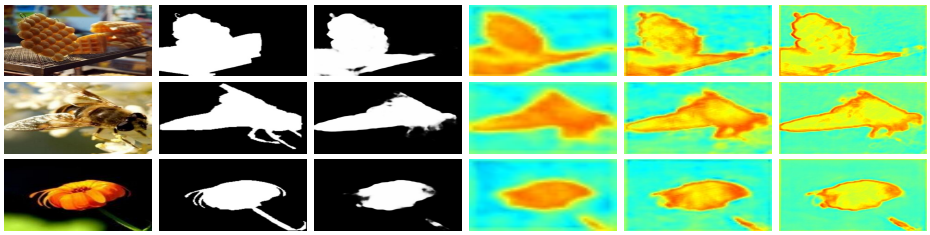
Depth Distillation (D) We test the effectiveness of depth distillation for DBD in Fig. 8(d)(e) and Table 2. It is clear that with the help of depth, our network can understand scene well and gain much better results because the depth information gives a strong prior for defocus map detection. Using Depth distillation, our network can also predict the relative depth from a single image. Although it is not our main target and our network can only predict the depth for partial defocus images, we still compare the distilled depth with our teacher network (Chen *et al.* [3]) in the supplementary materials.

Depth Distillation Hyper-Parameter γ We evaluate the influence of depth distillation hyper-parameters γ on DBD. Thus, we train our full method with different γ values. As shown in the Table 3, when γ is too large or too small, the performance become worse. Our network gain the best performance when γ equals to 0.1.

Selective Reception Field Block (SRFB) We evaluate the performance of the proposed SRFB by inserting the SRFB in each level of the decoder. As shown in Fig. 8 and Table 2, the SRFB models multi-scale features from the input and generate more accurate results. In addition to the necessity of our SRFB shown in Table 2 and Fig. 8, we also conduct the experiments to compare our SRFB

Table 3: Hyper-parameters γ for depth distillation. The best and second best results are marked in bold and underline, respectively.

Datasets	Metrics	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 5$
CUHK	F^β	<u>0.9253</u>	0.9208	**0.9267**	0.9231	0.9222
100	MAE	0.0438	0.0442	**0.0424**	0.0434	<u>0.0430</u>
DUT	F^β	0.8844	**0.8919**	<u>0.8909</u>	0.8902	0.8818
500	MAE	0.0737	0.0729	<u>0.0727</u>	**0.0696**	0.0786



(a) (b) (c) (d) (e) (f)

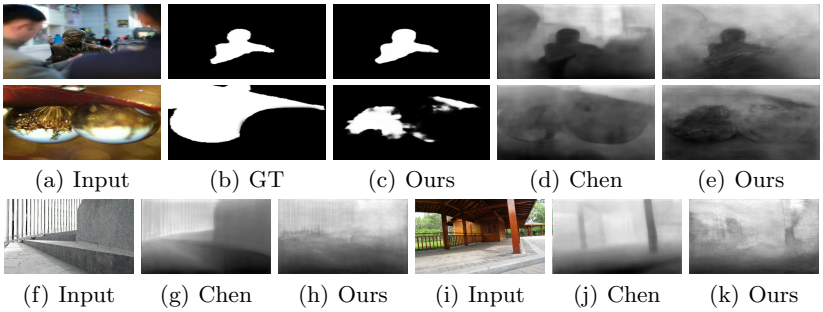
Fig. 9: Visualized attention maps in SAB. From left to right: (a) input, (b) target, (c) our final prediction, where (d)-(f) are the three attention maps in different levels of decoders. Here, we resize all the attentions to the same size for comparison. Interestingly, in high-level attention (d), our SAB generates the attention map for the whole defocus region, while in the coarser level (f), our attention map focuses on learning the edges and details.

with the model without selective attention (similar to FRB [22]). As shown in Fig. 8 and Table 2, although the FRB perform better in the CUHK100 dataset, our SRFB show a much better results in DUT500. We argue that CUHK100 is smaller and easier. Thus, the proposed SRFB is more suitable for DBD.

Supervision-guided Attention Block (SAB) In each level of auxiliary outputs, we design SAB to reuse the predicted defocus and depth map as spatial attention for further prediction. These attentions emphasize the useful features for further refinement. As shown in Fig. 8(c)(f) and Table 2, the proposed SAB also benefits blur detection. We also plot different levels of attention maps in the proposed SAB in Fig. 9. It is shown that using side outputs to generate the attention map emphasizes different features in each of their scales. Higher-level attentions stress the global features while the lower ones focus on local details.

4.3 Failure Cases

Although our network shows much better results than previous methods, there are still some failure cases. As shown in the first row of Fig. 10, when the far and near out-of-focus regions appears in a single image, the proposed network successfully predicts the defocus map but the relative depth relationship of the



(a) Input (b) GT (c) Ours (d) Chen (e) Ours

(f) Input (g) Chen (h) Ours (i) Input (j) Chen (k) Ours

Fig. 10: Failure cases. The top two rows show the failure examples when the scenes are complicated or when the depth is hard to predict. The thirds line shows the failed cases of predicted depth for all-in-focus image.

front person is incorrect. We also plot a more complicated example in the second row, the depth in this scene is hard to estimate because of the reflection of the water drop. Therefore, the proposed network cannot obtain global information and only predicts the scenes in the water drop. However, we argue that these problems can be mitigated by stronger networks and larger datasets.

Another limitation is our depth estimation. Our network can only predict the relative depth for partially defocused images, not depth estimation in the wild as in Chen *et al.* [3]. We randomly choose two all-in-focus images and plot the results in the third line of Fig. 10. When the image is all-in-focus, the defocus maps will not provide an effective prior for depth estimation. Thus, the apply range of our depth estimation is limited. However, our main target is DBD other than depth estimation.

5 Conclusions

In this paper, we firstly discuss the role of depth in defocus blur detection and propose depth distillation for this task. In detail, we distill the relative depth as regularization for learning-based defocus blur detection in a FCN network. Moreover, in order to build a stronger network, we design a selective reception field block because DBD is sensitive to multi-scale features, and we design a supervision-guided attention block, which serves the side outputs as spatial attention. The experimental results show the superiority of our method compared with 11 state-of-the-art methods in terms of efficiency and accuracy.

Acknowledgments

The authors would like to thanks Nan Chen for her helpful discussion. This work was partly supported by the University of Macau under Grants: MYRG2018-00035-FST and MYRG2019-00086-FST, and the Science and Technology Development Fund, Macau SAR (File no. 041/2017/A1, 0019/2019/A).

References

1. Bae, S., Durand, F.: Defocus magnification. In: Computer Graphics Forum. vol. 26, pp. 571–579. Wiley Online Library (2007)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE TPAMI **40**(4), 834–848 (2017)
3. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: NeurIPS. pp. 730–738 (2016)
4. Cun, X., Xu, F., Pun, C.M., Gao, H.: Depth-assisted full resolution network for single image-based view synthesis. IEEE computer graphics and applications **39**(2), 52–64 (2018)
5. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NeurIPS. pp. 2366–2374 (2014)
6. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR. pp. 270–279 (2017)
7. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 3828–3838 (2019)
8. Golestaneh, S.A., Karam, L.J.: Spatially-Varying Blur Detection Based on Multiscale Fused and Sorted Transform Coefficients of Gradient Magnitudes. CVPR (Mar 2017)
9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
10. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.: Deeply supervised salient object detection with short connections. In: CVPR. pp. 3203–3212 (2017)
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141 (2018)
12. Hu, X., Fu, C.W., Zhu, L., Heng, P.A.: Depth-attentional features for single-image rain removal. In: CVPR. pp. 8022–8031 (2019)
13. Hu, X., Zhu, L., Fu, C.W., Qin, J., Heng, P.A.: Direction-aware spatial context features for shadow detection. In: CVPR. pp. 7454–7462 (2018)
14. Jiang, P., Ling, H., Yu, J., Peng, J.: Salient region detection by ufo: Uniqueness, focusness and objectness. In: CVPR. pp. 1976–1983 (2013)
15. Karaali, A., Jung, C.R.: Image retargeting based on spatially varying defocus blur map. In: ICIP. pp. 2693–2697. IEEE (2016)
16. Karaali, A., Jung, C.R.: Edge-based defocus blur estimation with adaptive scale selection. IEEE TIP **27**(3), 1126–1137 (2017)
17. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NeurIPS. pp. 109–117 (2011)
18. Lee, J., Lee, S., Cho, S., Lee, S.: Deep defocus map estimation using domain adaptation. In: CVPR. pp. 12222–12230 (2019)
19. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: CVPR. pp. 510–519 (2019)
20. Li, Z., Dekel, T., Cole, F., Tucker, R., Snavely, N., Liu, C., Freeman, W.T.: Learning the depths of moving people by watching frozen people. In: CVPR. pp. 4521–4530 (2019)
21. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)

22. Liu, S., Huang, D., et al.: Receptive field block net for accurate and fast object detection. In: ECCV. pp. 385–400 (2018)
23. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (June 2015)
24. Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V.: Unifying distillation and privileged information. arXiv preprint arXiv:1511.03643 (2015)
25. Marc, L., Andrew, A., Nora, W.: Depth of field. http://graphics.stanford.edu/courses/cs178/applets/dof.html
26. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distill **1**(10), e3 (2016)
27. Park, J., Tai, Y.W., Cho, D., Kweon, I.S.: A Unified Approach of Multi-scale Deep and Hand-Crafted Features for Defocus Estimation. CVPR **cs.CV** (2017)
28. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: A benchmark and algorithms. In: ECCV. pp. 92–109. Springer (2014)
29. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointnets for 3d object detection from rgb-d data. In: CVPR. pp. 918–927 (2018)
30. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: CVPR (June 2019)
31. Qu, Y., Chen, Y., Huang, J., Xie, Y.: Enhanced pix2pix dehazing network. In: CVPR. pp. 8160–8168 (2019)
32. Shi, J., Li, X., Jia, J.: Discriminative Blur Detection Features. CVPR (2014)
33. Shi, J., Xu, L., Jia, J.: Just noticeable defocus blur detection and estimation. In: CVPR. pp. 657–665 (2015)
34. Tang, C., Zhu, X., Liu, X., Wang, L., Zomaya, A.: Defusionnet: Defocus blur detection via recurrently fusing and refining multi-scale deep features. In: CVPR. pp. 2700–2709 (2019)
35. Wikipedia contributors: Depth of field — Wikipedia, the free encyclopedia (2019), [Online; accessed 17-October-2019]
36. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: ECCV. pp. 3–19 (2018)
37. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: CVPR. pp. 3907–3916 (2019)
38. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K.: Aggregated Residual Transformations for Deep Neural Networks. CVPR (2017)
39. Xu, G., Quan, Y., Ji, H.: Estimating defocus blur via rank of local patches. In: CVPR. pp. 5371–5379 (2017)
40. Yi, X., Eramian, M.: Lbp-based segmentation of defocus blur. IEEE transactions on image processing **25**(4), 1626–1638 (2016)
41. Zhang, H., Patel, V.M.: Densely connected pyramid dehazing network. In: CVPR (2018)
42. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 2881–2890 (2017)
43. Zhao, W., Zhao, F., Wang, D., Lu, H.: Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network. In: CVPR. pp. 3080–3088 (2018)
44. Zhao, W., Zhao, F., Wang, D., Lu, H.: Defocus blur detection via multi-stream bottom-top-bottom network. IEEE TPAMI (2019)
45. Zhao, W., Zheng, B., Lin, Q., Lu, H.: Enhancing diversity of defocus blur detectors via cross-ensemble network. In: CVPR (June 2019)
46. Zheng, Q., Qiao, X., Cao, Y., Lau, R.W.: Distraction-aware shadow detection. In: CVPR (June 2019)

47. Zhu, L., Deng, Z., Hu, X., Fu, C.W., Xu, X., Qin, J., Heng, P.A.: Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In: ECCV (2018)