This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Enhancing Diversity of Defocus Blur Detectors via Cross-Ensemble Network

Wenda Zhao, Bowen Zheng, Qiuhua Lin, Huchuan Lu
Dalian University of Technology, China
zhaowenda@dlut.edu.cn, sltk@mail.dlut.edu.cn, {qhlin, lhchuan}@dlut.edu.cn

Abstract

Defocus blur detection (DBD) is a fundamental yet challenging topic, since the homogeneous region is obscure and the transition from the focused area to the unfocused region is gradual. Recent DBD methods make progress through exploring deeper or wider networks with the expense of high memory and computation. In this paper, we propose a novel learning strategy by breaking DBD problem into multiple smaller defocus blur detectors and thus estimate errors can cancel out each other. Our focus is the diversity enhancement via cross-ensemble network. Specifically, we design an end-to-end network composed of two logical parts: feature extractor network (FENet) and defocus blur detector cross-ensemble network (DBD-CENet). FENet is constructed to extract low-level features. Then the features are fed into DBD-CENet containing two parallel-branches for learning two groups of defocus blur detectors. For each individual, we design cross-negative and self-negative correlations and an error function to enhance ensemble diversity and balance individual accuracy. Finally, the multiple defocus blur detectors are combined with a uniformly weighted average to obtain the final DBD map. Experimental results indicate the superiority of our method in terms of accuracy and speed when compared with several state-of-the-art methods.

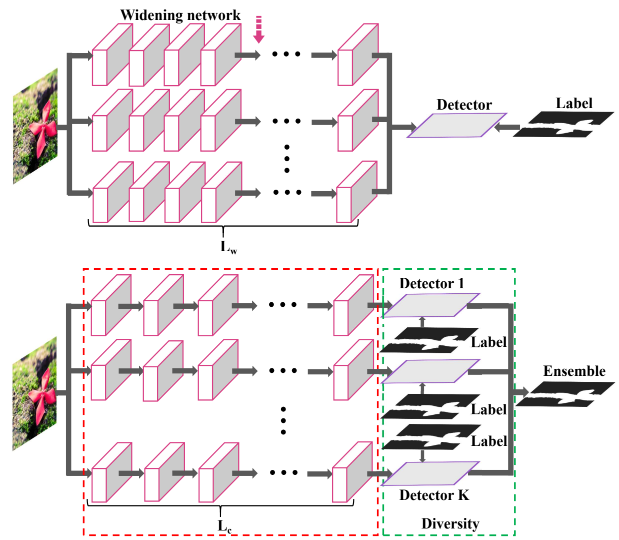


Figure 1. Motivation for our method. The first row expresses a representative model (*e.g.*, [33]) with wider and deeper network to achieve progress for DBD. However, one detector lacks diversity. We adopt the division of labor method by breaking it into multiple smaller defocus blur detectors (the number of convolutional layer $L_c < L_w$). Our concern is mainly how to enhance diversity of these defocus blur detectors (green rectangle box) and how to implement network with less parameters (red rectangle box).

1. Introduction

Defocus blur detection (DBD) is a fundamental topic in computer vision. The goal is to highlight the defocus blur or focus region in an image. DBD is of broad interest for potential applications of autofocus [32], depth detection [19], image retargeting [10], blur reconstruction [26], *etc*.

Conventional DBD methods usually utilize hand-crafted low-level features such as gradient [5, 11, 27], contrast [28] and frequency [22, 25] to distinguish defocus blur. However, it is of great difficulty for these low-level features-based methods to discriminate defocus blur regions in complex scenarios, especially in the presence of the homogeneous regions (*e.g.*, such regions show almost no difference in appearance when they are in-focus or out-of-focus) and the obscure boundaries between focused and unfocused areas.

Recently, Deep Convolutional Neural Networks (DCNNs), which intelligently learn hierarchical representation of the input directly, have achieved superior performance in many vision tasks. There have been fairly large efforts to enhance DCNNs with greater capacity, *e.g.*, increased depth [24, 6], enlarged width [29, 2, 7], novel layers [30, 8, 13], *etc*. Inspired by this, several DCNNs-based DBD models have been proposed. [16, 9] measure defocus blur in a patch-by-patch scanning manner, leading to numerous redundant computations. Despite the progress for DBD has been achieved later by exploring deeper or wider networks [33, 31], their single detectors lack diversity and large number of parameters cause large computation cost.

In this paper, we contribute from a different view (see Figure 1), and propose a deep defocus blur detector cross-ensemble network (CENet). We adopt the division of labor method, simplifying the DBD problem by breaking it into multiple smaller defocus blur detectors. Then, these detectors are combined with a uniformly weighted average, hopefully reducing the DBD error compared to a single defocus blur detector. For each individual, we design cross-negative and self-negative correlations and an error function to enhance ensemble diversity and balance individual accuracy.

The core idea of CENet is to enhance diversity of defocus blur detectors via cross-ensemble network, and thus estimate errors may cancel out each other. Our focus is the diversity enhancement with cross-ensemble strategy and computation efficiency. Therefore, we consider an end-to-end DCNN to be composed of two logical parts: feature extractor network (FENet) and defocus blur detector cross-ensemble network (DBD-CENet). FENet is designed to extract low-level features. Then the features are fed into DBD-CENet consisting of two subnetworks side by side to learn two groups of defocus blur detectors. Each detector of the current group is alternately optimized by cross-negative and self-negative correlation losses and an error function to penalize the correlation with the other group and the current group to enhance diversity. We implement multiple shallow networks to produce defocus blur detectors. In addition, we adopt convolutional feature-shared strategy to further reduce network parameters.

To sum up, our main contributions are three-fold:

- We present a novel perspective for DBD by enhancing diversity of defocus blur detectors with cross-ensemble network. Two groups of defocus blur detectors are alternately optimized through respective cross-negative and self-negative correlations to enhance diversity.

- Multiple shallow networks are utilized to produce defocus blur detectors, and convolutional feature-shared strategy is adopted to implement CENet. Compared to the deeper or wider network (*e.g.*, [33]), CENet is practically feasible with less computation cost.

- Extensive performance evaluation indicates that our model outperforms the other state-of-the-art methods in terms of DBD accuracy and calculating speed.

2. Related Work

Traditional methods. DBD is a basic issue of computer vision where adopted features are critical for the detection performance. Most of traditional DBD methods are based on low-level manually designed features [11, 20, 22]. For example, Golestaneh *et al.* [5] use high-frequency multi-scale fusion and sort transform of gradient magnitudes to compute blur detection maps. Xu *et al.* [27] present a metric for DBD at edge points through the maximum ranks of the corresponding local patches with different orientations in gradient domain. Yi *et al.* [28] adopt local binary patterns for focus sharpness metric. Without high-level semantic information, the hand-crafted feature-based methods usually cause inaccurate detection especially in the presence of the homogeneous areas.

DCNN methods. Recent works [16, 9, 33, 31] have resorted to the deep convolutional neural networks, which have set new state-of-the-art on DBD. On the one hand, Huang *et al.* [9] design a patch-level CNN to learn discriminative deep blur features. Park *et al.* [16] combine deep patch-level and hand-crafted features to fed into a fully connected neural network classifier to estimate the degree of defocus. Patch-level DCNN methods is very time consuming, which is needed to run thousands of times to process a raw image. On the other hand, making the network go deeper or wider has also been proven applicable. Zhao *et al.* [33] propose a multi-stream bottom-top-bottom fully convolutional network that handles input images with different scales, and then design a fusion and recurrent reconstruction strategy to widen and deepen network to improve the performance of DBD. Zhang *et al.* [31] use a dilated fully convolutional neural network with pyramid pooling and boundary refinement layers to widen network for generating good DBD map. Despite the improvement for DBD has been achieved, their large number of parameters lead to high storage and computation consumption. Different from these previous works, focusing to enhance diversity, we design a CENet to obtain multiple defocus blur detectors with less learning parameters, hopefully reducing the DBD error and increasing computing speed.

Ensemble methods. Ensemble learning is a widely used and efficacious technique in machine learning, and its success is commonly attributed to the diversity within the ensemble [14, 18, 1, 17]. Recently, similar ideas have been applied to DCNN methods and improve performance for various visual tasks. For example, Pan *et al.* [15] propose a DualCNN for low-level vision, which consists of two parallel branches to respectively obtain structure detector and detail detector. However, this strategy relies heavily on manual structure and detail labels, which limits the diversity. Hou *et al.* [7] use two parallel neural networks to automatically learn diversity for image recognition. But it is not clear how these two networks can alternatively assist each other. [23] learns a pool of decorrelated regressors for crowd counting. However, a pool of decorrelated regressors are not enough to generate adequate diversity for our DBD task. In contrast, our CENet learns two groups of defocus blur detectors, which is alternately optimized with cross-negative and self-negative correlation losses to enhance diversity.

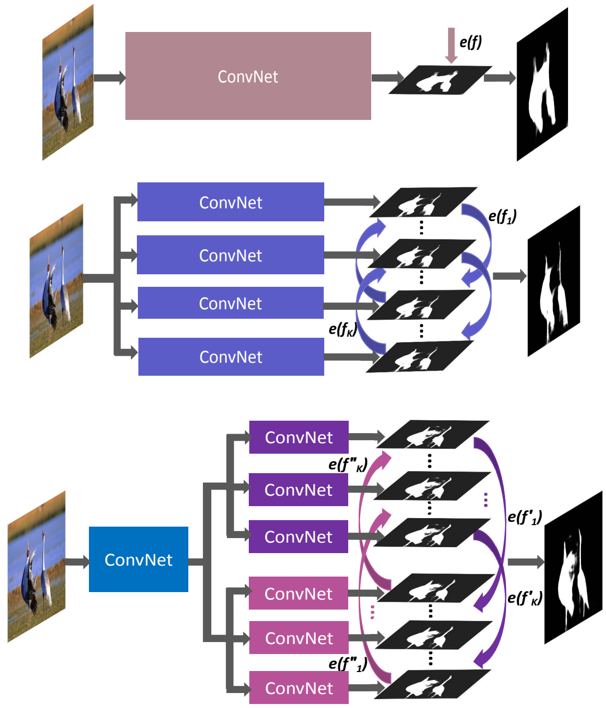


Figure 2. Different detector ensemble methods. From top row to bottom row are SENet, MENet and CENet, respectively. *ConvNet* stands for a cascade convolutional layers and pooling layers, and each layer is composed of a convolution and a rectified linear unit. *ConvNets* with the same color express for sharing parameters. Architecture details refer to Subsection 3.2.

3. Deep Cross-Ensemble Network

3.1. Formulation

We assume that we have access to N training samples, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$. The samples are M dimensional, $M = H \times W \times C$, where H, W and C denote the height, width and number of channels of the i-th sample, respectively. We learn to estimate a DBD map $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$, where each $\mathbf{y}_i$ in D dimensional space, $D = H \times W$. The learning problem is to use the set $\mathbf{X}$ to learn a detector or multiple detectors to approximate the correct mapping from input to output. With different number of detectors, the network models are divided into three categories: single detector network (SENet), multi-detector ensemble network (MENet) and cross-ensemble network (CENet). Their formulations are explained below.

SENet. Single detector-based networks are accustomed for DBD tasks, *e.g.,* [33, 31]. In order to improve performance, deepening or widening the network is a common strategy. However, numerous parameters make the optimization difficult and computation consumable. Here, we use a parameterized detector f to find the set of parameters $\mathbf{w}$ that minimise the expected mean squared error

$$e(f) = \frac{1}{N}\sum_{i=1}^{N}(f(\mathbf{x}_i;\mathbf{w}) - \mathbf{y}_i)^2, \tag{1}$$

where $e(\cdot)$ indicates mean squared error. SENet framework is shown in Figure 2. Due to the lack of diversity, SENet can hardly achieve optimal results (see Figure 3).

MENet. Instead of a single detector f, MENet contains a group of them: $F = \{f_1, f_2, ..., f_K\}$, where each f_k has its own parameter vector $\mathbf{w}_k$, and K is the total number of detectors. We can train a group of detectors jointly. Once this is accomplished, the outputs of the individuals are combined to obtain the ensemble $\widehat{f}$. Here, we take a uniformly weighted average as the combination mechanism

$$\widehat{f}(\mathbf{X};\mathbf{w}_1,\mathbf{w}_2,...,\mathbf{w}_K) = \frac{1}{K}\sum_{k=1}^{K} f_k(\mathbf{X};\mathbf{w}_k). \tag{2}$$

We adopt the ensemble method [14], which has shown a number of empirical successes for many applications [23, 21]. Treating the ensemble $\widehat{f}$ as a single learning unit, [4], [14], [1] use the bias-variance decomposition as

$$E\{(\widehat{f} - \mathbf{Y})^2\} = (E\{\widehat{f}\} - \mathbf{Y})^2 + E\{(\widehat{f} - E\{\widehat{f}\})^2\}, \tag{3}$$

where the shorthand expectation operator $E\{\cdot\}$ is used to represent the generalization ability. Then, given Equation (2), it can show

$$\begin{aligned} E\{(\widehat{f} - \mathbf{Y})^2\} =& \frac{1}{K^2}(\sum_{k=1}^{K}(E\{f_k\} - \mathbf{Y}))^2 \\ &+ \frac{1}{K^2}\sum_{k=1}^{K}E\{(f_k - E\{f_k\})^2\} \\ &+ \frac{1}{K^2}\sum_{k=1}^{K}\sum_{j\neq k}E\{(f_k - E\{f_k\})(f_j - E\{f_j\})\}. \end{aligned} \tag{4}$$

After some arrangements, it has

$$\begin{aligned} E\{(\widehat{f} - \mathbf{Y})^2\} =& \frac{1}{K}(\sum_{k=1}^{K}E\{(f_k - \mathbf{Y})^2\} \\ &- \frac{1}{K}\sum_{k=1}^{K}E\{(f_k - \widehat{f})^2\}, \end{aligned} \tag{5}$$

where the first item is the weighted average error of the individuals, and the second item measures the amount of correlation between the ensemble and each individual. Based on this, we train a group of detectors with each objective loss

$$e(f_k) = \frac{1}{2}(f_k(\mathbf{X};\mathbf{w}_k) - \mathbf{Y})^2 - \lambda(f_k(\mathbf{X};\mathbf{w}_k) - \widehat{f})^2, \tag{6}$$

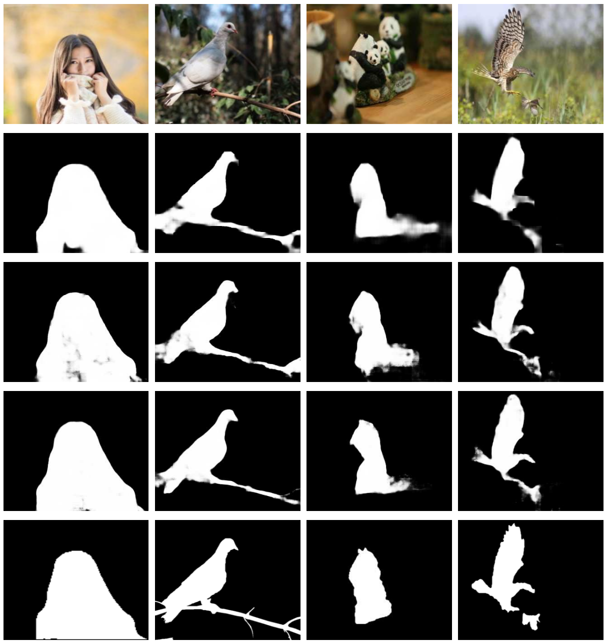


Figure 3. Comparison of DBD maps generated by our proposed networks. From top row to bottom row are source images, SENet results, MENet results, CENet results and ground truthes, respectively. It can be seen from that CENet consistently produces DBD maps closest to the ground truth.

where non-negative weight λ expresses the trade-off between these two items. The second term in Eq. 6 penalizes the correlation of each detector with others to make better trade-offs among the accuracy and diversity for reducing the overall loss function. MENet framework is shown in Figure 2. Figure 3 demonstrates that MENet achieves better performance than SENet. However, MENet can not generate adequate diversity for our DBD task.

CENet. Although MENet improves the detection accuracy, it has limit when the input image has small-scale focused area or large-scale homogeneous regions (see Figure 3). The main reason is that MENet does not effectively encourage these detectors diversity. Thus, we further propose a CENet, which is constructed with two groups of defocus blur detectors, $F' = \{f'_1, f'_2, ..., f'_K\}$ and $F'' = \{f''_1, f''_2, ..., f''_K\}$. Each group of detectors has their own parameters $\mathbf{W}' = \{\mathbf{w}'_1, \mathbf{w}'_2, ..., \mathbf{w}'_K\}$ and $\mathbf{W}'' = \{\mathbf{w}''_1, \mathbf{w}''_2, ..., \mathbf{w}''_K\}$, respectively. Consider Equation (6) from another perspective, we can divide it into two items: an error function to balance individual accuracy and a negative correlation to achieve ensemble diversity. To enhance ensemble diversity, we further propose a cross-negative correlation loss. Each detector is not only negatively correlated with the other detectors of the current group, but also with the ones of the other one group. For individual detector in the first group of detectors, we have

$$
\begin{aligned}
e(f'_k) = &\frac{1}{2}(f'_k(\mathbf{X};\mathbf{w}'_k) - \mathbf{Y})^2 \\
&- \lambda(f'_k(\mathbf{X};\mathbf{w}'_k) - \widehat{f'})^2 - \gamma(f'_k(\mathbf{X};\mathbf{w}'_k) - \widehat{f''})^2,
\end{aligned} \tag{7}
$$

where $\widehat{f''}$ is the uniformly weighted average of the second group of detectors. The first item is to assure accuracy; the second item aims to enhance diversity of the the current group; the third item focuses on improving diversity with the other one group. λ and γ stands for non-negative trade-off weights. The loss of each detector in the second group is as follows

$$
\begin{aligned}
e(f''_k) = &\frac{1}{2}(f''_k(\mathbf{X};\mathbf{w}''_k) - \mathbf{Y})^2 \\
&- \lambda(f''_k(\mathbf{X};\mathbf{w}''_k) - \widehat{f''})^2 - \gamma(f''_k(\mathbf{X};\mathbf{w}''_k) - \widehat{f'})^2,
\end{aligned} \tag{8}
$$

where $\widehat{f'}$ illustrates the uniformly weighted average of the first group of detectors. The two groups of detectors are alternately optimized to enhance diversity. CENet framework is shown in Figure 2. In Figure 3, DBD maps generated by CENet highlight focused area the most uniformly and have the sharpest boundaries on the transition region. Objective comparison is shown in Subsection 4.3.

3.2. Architecture Details

In this work, we propose to enhance diversity of defocus blur detectors via cross-ensemble network. Our focus is the diversity enhancement with cross-ensemble strategy and computation efficiency. To show the effectiveness of cross-ensemble strategy, we resort to plain CNN without complex architectural engineering. We employs VGG16 [24] network and make several modifications as the baseline.

Figure 4 shows the architecture of the proposed CENet, which includes two networks, FENet and DBD-CENet. Trivially applying ensemble learning to train multiple convolutional networks will significantly increase memory cost. Instead, we adopt the division of labor method and convolutional feature-shared strategy to reduce memory cost. Thus, we design FENet to extract low-level features, which are shared by the following parallel detector branches. FENet is constructed with the first two convolutional blocks (CB_1 and CB_2) of VGG16 acting as a feature extractor, which takes the original RGB images as input and produces low-level feature maps of 128 channels. The feature maps from FENet are then fed into the following DBD-CENet consisting of parallel defocus blur detector branches. The networks of each branche for producing defocus blur detectors share parameters to reduce parameters. Therefore, each branche consists of last three fully convolutional blocks of CB_3, CB_4 and CB_5 (or CB'_3, CB'_4 and CB'_5) of VGG16, and is followed by a convolution layer

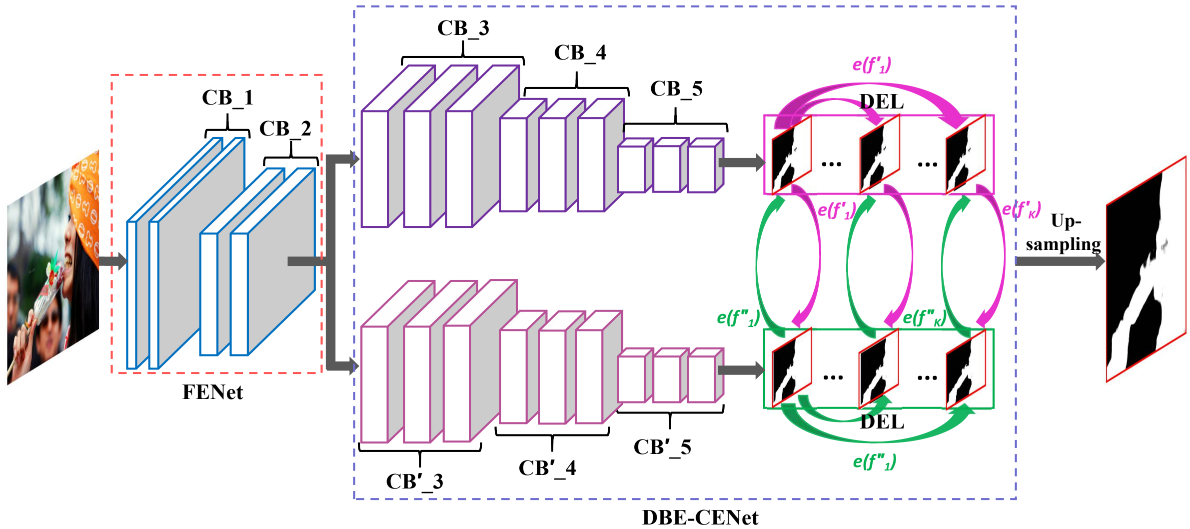


Figure 4. The architecture of our proposed CENet. Each colorful box is considered as a convolutional block (CB). Each convolutional block followed by a pooling layer except for the last one. We use VGG16 [24] and make several modifications as the baseline. The input RGB image is processed by FENet, which consists of two convolutional blocks: CB_1 and CB_2, to extract low-level features. Then the features are fed into DBD-CENet containing two parallel-branches for learning two groups of diversity defocus blur detectors. The networks of each group for producing defocus blur detectors share parameters. Therefore, each branche is composed of three convolutional blocks: CB_3, CB_4 and CB_5 (or CB'_3, CB'_4 and CB'_5), and the top is a detector generation layer (DGL) consisting of a convolution layer with K channels, respectively. Finally, the multiple defocus blur detectors are combined and followed by a up-sampling operation to obtain the final DBD map.

with K channels to convert 512 channels to K DBD detectors. Hence, we name this convolution block as detector generation layer (DGL). Finally, the DBD map will be obtained by combining the defocus blur detectors produced by the two branches.

SENet and MENet include FENet and one branch of DBD-CENet. To achieve fair comparison with CENet, SENet and MENet are designed for the same capacity as CENet. Specifically, the channels of each convolutional layers in last three fully convolutional blocks of VGG16 are doubled.

3.3. Training

We adopt iterative training strategy to train CENet. We first train FENet and one branch of DBD-CENet with pre-trained parameters of VGG16 network on ImageNet [3]. Then we fix FENet and initialize the other branch with the trained parameters of the first branch. Finally, we finetune the two branches of DBD-CENet in an iterative way with every epoch. The gradients for two defocus blur detector branches can be respectively obtained by

$$\frac{\partial e(f'_k)}{\partial f'_k} = (f'_k(\mathbf{X};\mathbf{w}'_k) - \mathbf{Y}) - 2\lambda(f'_k(\mathbf{X};\mathbf{w}'_k) - \widehat{f'}) - 2\gamma(f'_k(\mathbf{X};\mathbf{w}'_k) - \widehat{f''}), \tag{9}$$

Algorithm 1: Training algorithm for CENet.

Input: Training data $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$.
Initialize:
1. Train FENet and F'.
2. Fix FENet and initialize F'' with parameters of F'.
3. **while** training is not convergent **do**
4. Calculate $\widehat{f''} = \frac{1}{K}\sum_{k=1}^{K} f''_k$.
5. Update defocus blur detectors $\{f'_1, f'_2, ..., f'_K\}$ using Equation (9).
6. Calculate $\widehat{f'} = \frac{1}{K}\sum_{k=1}^{K} f'_k$.
7. Update defocus blur detectors $\{f''_1, f''_2, ..., f''_K\}$ using Equation (10).
8. **end while**
Output: F' and F''.

$$\frac{\partial e(f''_k)}{\partial f''_k} = (f''_k(\mathbf{X};\mathbf{w}''_k) - \mathbf{Y}) - 2\lambda(f''_k(\mathbf{X};\mathbf{w}''_k) - \widehat{f''}) - 2\gamma(f''_k(\mathbf{X};\mathbf{w}''_k) - \widehat{f'}). \tag{10}$$

Algorithm 1 illustrates the training procedure of the proposed CENet.

In the test stage, we compute the final DBD map using

Table 2. Comparing CENet with SENet and MENet using F-measure and MAE values on both DUT and CUHK datasets.

Method	DUT		CUHK	
	F-measure	MAE	F-measure	MAE
SENet	0.750	0.152	0.884	0.066
MENet	0.758	0.149	0.896	0.062
CENet	**0.789**	**0.135**	**0.906**	**0.059**

the two groups of detectors according to the formation

$$Y = \frac{1}{K}\sum_{k=1}^{K} f'_k(X;\mathbf{w}'_k) + \frac{1}{K}\sum_{k=1}^{K} f''_k(X;\mathbf{w}''_k), \tag{11}$$

where Y stands for the final DBD map and X illustrates the input image.

4. Experiments

4.1. Experimental Setup

Benchmark datasets. Two publicly available benchmark datasets with pixel-wise annotations are used in this work. The first one is the CUHK blur dataset [22], which consists of 704 partially defocus blurred images. It includes, but not limited to, various scenes with cluttered backgrounds. The second benchmark dataset is the DUT defocus blur dataset [33]. It includes 500 test images, and numerous images contain multi-scale focused areas. We adopt the strategy of [33] to train our model where 604 images of CUHK blur dataset for training and the remaining 100 images and DUT dataset for testing. Also, data augmentation (e.g., flipping) is performed.

Implementation details. We implement the proposed network and training procedures in TensorFlow with a GTX 1080TI GPU. Adam [12] is used to optimize our network with the momentum value of 0.9, and a weight decay of 5e-3. We use a fixed learning rate 1e-4 with mini-batch of 2. The code is available at https://github.com/wdzhao123/CENet-code.

Evaluation metrics. We use three evaluation metrics, including the mean absolute error (MAE), F-measure, and the Precision-Recall (PR) curve. MAE can provide a good measure of the dissimilarity between the ground truth and DBD map. It is the average per-pixel difference between the ground truth and the DBD map, which is defined as:

$$MAE = \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H} \left| Y(x,y) - G(x,y) \right|, \tag{12}$$

where x, y stands for pixel coordinates. W and H denotes the width and height of the DBD map Y, respectively, and G is the ground truth.

The F-measure, which is an overall performance measurement, is defined as:

$$F_\beta = \frac{(1+\beta^2)\cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}, \tag{13}$$

where $\beta^2 = 0.3$ is employed to emphasize the precision. $Precision$ stands for the percentage of focused pixels being correctly detected, and $Recall$ is the fraction of detected focused pixels in relation to the ground truth number of focused pixels. Each map is binarized with an adaptive threshold, which is 1.5 times the mean value of the DBD map.

4.2. Comparison with the State of the Art

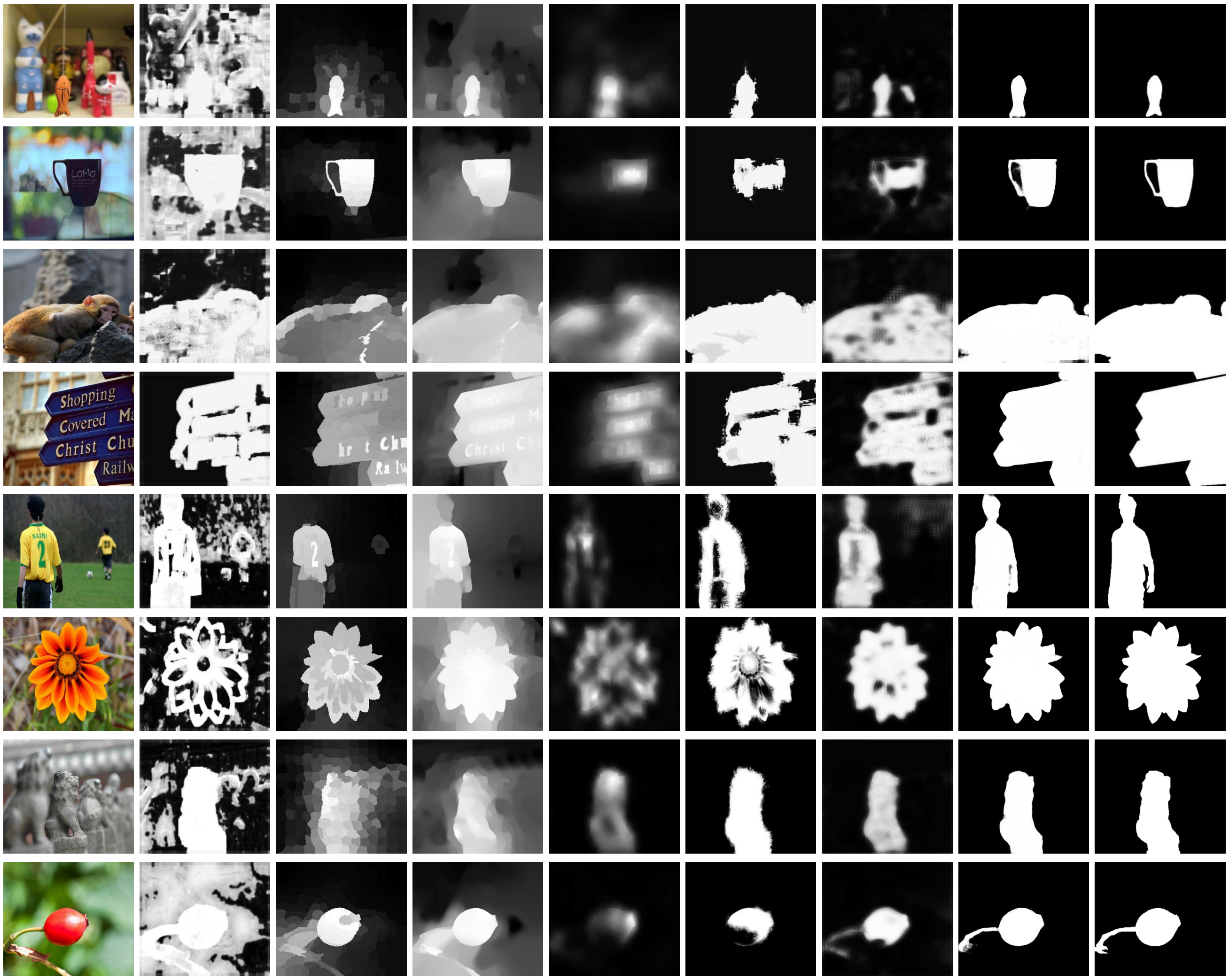
Our method is compared with six state-of-the-art DBD methods, including discriminative blur detection features (DBDF) [22], spectral and spatial approach (SS) [25], deep and hand-crafted features (DHCF) [16], high-frequency multi-scale fusion and sort transform of gradient magnitudes (HiFST) [5], local binary patterns (LBP) [28] and multi-stream bottom-top-bottom fully convolutional network (BTBNet) [33]. For fair comparison, we utilize the recommended parameter settings to implement these methods or adopt the DBD maps provided by authors.

We provide visual comparison results in Figure 5, which show various challenging cases, *e.g.*, a wide variety of scenes with cluttered backgrounds and multi-scale focused areas from small scale to big scale. It can be seen that our CENet highlights focused area the most uniformly and produces the sharpest boundaries on the transition region. Our method has the best results for different scale focused area detection (*e.g.*, the small scale and large scale ones in the second and third rows of Figure 5). In the fifth row of Figure 5, almost all methods produce noise because of the cluttered background and homogeneous region except for our method.

Table 1 reports the comparison results, where we can see that our method outperforms the others in terms of both F-measur and MAE for both benchmark datasets. Comparing F-measur scores, our CENet outperforms the second best method by 2.9% and 5.2% over DUT and CUHK respectively. Moreover, our method lowers the MAE scores significantly on both datasets. Meanwhile, our CENet is also highly efficient with the speed of 15.63 FPS, which is 6.2 times faster than the second fastest method SS [25]. PR curves and F-measure scores are displayed in Figure 6 and Figure 7, respectively. We can see that our proposed method performs favorably against other methods on both datasets.

4.3. Ablation Analysis

Comparison with SENet and MENet. To verify the va-

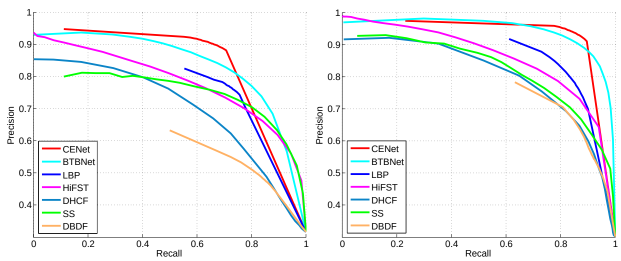


(a) Source (b) DBDF [22] (c) SS [25] (d) DHCF [16] (e) HiFST [5] (f) LBP [28] (g) BTBNet [33] (h) CENet (i) GT

Figure 5. Visual comparison of DBD maps generated by the proposed method and other state-of-the-art ones. The ground truth is shown in the last column. The first four sources are selected from the DUT dataset. The last four sources are chosen from CUHK dataset. It can be seen from that our model consistently produces DBD maps closest to the ground truth.

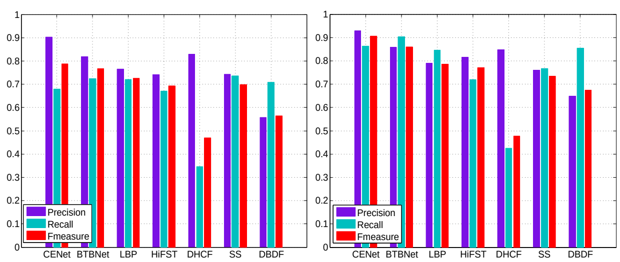
Table 1. Quantitative comparison of F-measure and MAE scores. The speeds of all methods tested on a workstation with an Intel 3.4GHz CPU with 32G memory for 320×320 images are listed in the last row. The best two results are shown in **red** and **green** colors, respectively.

	Metric	DBDF [22]	SS [25]	DHCF [16]	HiFST [5]	LBP [28]	BTBNet [33]	CENet
DUT	F-measure	0.565	0.699	0.470	0.693	0.726	**0.767**	**0.789**
	MAE	0.379	0.289	0.408	0.246	**0.191**	0.192	**0.135**
CUHK	F-measure	0.674	0.734	0.477	0.770	0.786	**0.861**	**0.906**
	MAE	0.289	0.229	0.372	0.220	0.136	**0.111**	**0.059**
DUT & CUHK	FPS	0.022	**2.530**	0.085	0.021	0.111	0.040	**15.63**



(a) DUT dataset (b) CUHK dataset

Figure 6. Comparison of precision-recall curves of the proposed method with other state-of-the-art methods using (a) DUT dataset and (b) CUHK dataset.



(a) DUT dataset (b) CUHK dataset

Figure 7. Comparison of the average precision, recall, and F-measure scores across (a) DUT dataset and (b) CUHK dataset. The proposed method achieves the highest F-measure on both datasets.

lidity of enhancing diversity of defocus blur detectors, we compare our model with SENet and MENet which are conducted for the same capacity as CENet (see Subsection 3.2). Table 2 reports the comparison results, showing that our cross-ensemble network enhances diversity of defocus blur detectors thus achieve better performance.

Effectiveness of parameters γ and λ. As described in Subsection 3.1, γ and λ in Equation (7) and Equation (8) stand for the trade-off between the cross-negative correlation and self-negative correlation. A larger γ will encourage cross diversity and a larger λ will encourage self diversity. But too large γ and λ will reduce individual accuracy. We first study the effect of γ and set λ to 0. It can be seen from Table 3 that $\gamma = 0.01$ can produce the better results. Then, we set γ to 0.01, and adjust the parameter λ. In Table 3, we can see that $\lambda = 0.1$ achieves the best results on both datasets.

Selection of parameter K. Parameter K expresses for the number of defocus blur detectors in the cross-ensemble network. In Table 4, we report the comparison results of CENet with different values of K. As can be seen that more detectors can achieve better performance. Considering model complexity and computational efficiency, we take K to 64, which has achieved the state of the art.

Table 3. Effect of parameters γ and λ using F-measure and MAE values on both DUT and CUHK datasets.

Method	DUT		CUHK	
	F-measure	MAE	F-measure	MAE
Our CENet with different values of γ ($\lambda = 0$)				
CENet with $\gamma = 0.1$	0.740	0.160	0.899	0.063
CENet with $\gamma = 0.05$	0.746	0.157	0.898	0.061
CENet with $\gamma = 0.01$	0.766	**0.144**	0.904	**0.059**
CENet with $\gamma = 0.005$	**0.770**	0.146	**0.905**	0.062
CENet with $\gamma = 0.001$	0.762	0.146	0.901	0.062
Our CENet with different values of λ ($\gamma = 0.01$)				
CENet with $\lambda = 0.1$	**0.789**	**0.135**	**0.906**	**0.059**
CENet with $\lambda = 0.05$	0.781	0.137	**0.906**	0.063
CENet with $\lambda = 0.01$	0.783	0.139	0.903	0.061
CENet with $\lambda = 0.005$	0.781	0.138	0.904	0.062
CENet with $\lambda = 0.001$	0.778	0.140	0.905	0.060

Table 4. Comparison of CENet with different values of K using F-measure and MAE scores on both DUT and CUHK datasets.

Method	DUT		CUHK	
	F-measure	MAE	F-measure	MAE
CENet with $K = 16$	0.786	0.136	0.897	0.069
CENet with $K = 32$	**0.793**	**0.135**	0.899	0.069
CENet with $K = 64$	0.789	**0.135**	**0.906**	**0.059**

5. Conclusions

In this paper, we present an effective cross-ensemble network to enhance diversity of defocus blur detectors. We pose DBD as an ensemble learning problem and learn two groups of defocus blur detectors. Each detector of the current group is alternately optimized by the cross-negative correlation and self-negative correlation losses to penalize the correlation with the other group and the current group to enhance diversity. Besides, we add an error function for each detector to balance individual accuracy. The core of our method is the introduction of cross-ensemble strategy, which aims to achieve the diversity enhancement of defocus blur detectors. In addition, we adopt feature-shared strategy to implement CENet, which is practically feasible with less number of parameters and computation consumption. Extensive experimental results on existing datasets verify that our method outperforms the state-of-the-art methods in terms of DBD accuracy and calculating speed.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant Nos. 61801077, 61872056, 61829102, 61871067, 61771088, 61725202, 61751212, the China Postdoctoral Science Foundation under Grant No. 2017M611221, and the Fundamental Research Funds for the Central Universities under Grant Nos. DUT16RC(3)077.

References

[1] G. Brown, J. L. Wyatt, and P. Tino. Managing diversity in regression ensembles. *Journal of Machine Learning Research*, 6:1621–1650, 2005. 2, 3

[2] C. Cheng, Y. Fu, W. Lu, Y. . G. Jiang, J. Feng, and X. Xue. Left-right skip-densenets for coarse-to-fine object categorization. *arXiv:1710.10386*, pages 1–9, 2017. 1

[3] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F.-F. Li. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[4] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992. 3

[5] S. A. Golestaneh and L. J. Karam. Spatially-varying blur detection based on multiscale fused and sorted transform coefficients of gradient magnitudes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 596–605, 2017. 1, 2, 6, 7

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[7] S. Hou, X. Liu, and Z. Wang. Dualnet: Learn complementary features for image recognition. In *IEEE International Conference on Computer Vision*, pages 502–510, 2018. 1, 2

[8] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv:1709.01507*, pages 1–11, 2017. 1

[9] R. Huang, W. Feng, M. Fan, L. Wan, and J. Sun. Multiscale blur detection by learning discriminative deep features. *Neurocomputing*, 285:154–166, 2018. 1, 2

[10] A. Karaali and C. R. Jung. Image retargeting based on spatially varying defocus blur map. In *IEEE International Conference on Image Processing*, pages 2693–2697, 2016. 1

[11] A. Karaali and C. R. Jung. Edge-based defocus blur estimation with adaptive scale selection. *IEEE Transactions on Image Processing*, 27(3):1126–1137, 2018. 1, 2

[12] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, pages 1–14, 2014. 6

[13] P. Li, J. Xie, Q. Wang, and Z. Gao. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. *arXiv:1712.01034*, pages 1–9, 2017. 1

[14] Y. Liu and X.Yao. Ensemble learning via negative correlation. *Neural Networks*, 12(10):1399–1404, 1999. 2, 3

[15] J. Pan, S. Liu, D. Sun, J. Zhang, Y. Liu, J. S. J. Ren, Z. Li, J. Tang, H. Lu, Y. Tai, and M. Yang. Learning dual convolutional neural networks for low-level vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3070–3079, 2018. 2

[16] J. Park, Y. Tai, D. Cho, and I. S. Kweon. A unified approach of multi-scale deep and hand-crafted features for defocus estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2760–2769, 2017. 1, 2, 6, 7

[17] Y. Ren, L. Zhang, and P. N. Suganthan. Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational Intelligence Magazine*, 11:41–53, 2016. 2

[18] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1619–1630, 2006. 2

[19] P. Sakurikar and P. J. Narayanan. Composite focus measure for high quality depth maps. In *IEEE International Conference on Computer Vision*, pages 1623–1631, 2017. 1

[20] P. Sakurikar and P. J. Narayanan. Composite focus measure for high quality depth maps. In *IEEE International Conference on Computer Vision*, pages 1623–1631, 2017. 2

[21] W. Sheng, P. Shan, S. Chen, Y. Liu, and F. E. Alsaadi. A niching evolutionary algorithm with adaptive negative correlation learning for neural network ensemble. *Neurocomputing*, 247:173–182, 2017. 3

[22] J. Shi, L. Xu, and J. Jia. Discriminative blur detection features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2965–2972, 2014. 1, 2, 6, 7

[23] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M.-M. Cheng, and G. Zheng. Crowd counting with deep negative correlation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5382–5390, 2018. 2, 3

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, pages 1–14, 2014. 1, 4, 5

[25] C. Tang, J. Wu, Y. Hou, P. Wang, and W. Li. A spectral and spatial approach of coarse-to-fine blurred image region detection. *IEEE Signal Processing Letters*, 23(11):1652–1656, 2016. 1, 6, 7

[26] Y. Wang, Z. Wang, D. Tao, S. Zhuo, X. Xu, S. Pu, and M. Song. Allfocus: Patch-based video out-of-focus blur reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(9):1895–1908, 2017. 1

[27] G. Xu, Y. Quan, and H. Ji. Estimating defocus blur via rank of local patches. In *IEEE International Conference on Computer Vision*, pages 5381–5389, 2017. 1, 2

[28] X. Yi and M. Eramian. LBP-based segmentation of defocus blur. *IEEE Transactions on Image Processing*, 25(4):1626–1638, 2016. 1, 2, 6, 7

[29] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv:1605.07146*, pages 1–15, 2016. 1

[30] H. Zhang, K. J. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018. 1

[31] S. Zhang, X. Shen, Z. Lin, R. Mech, J. P. Costeira, and J. M. F. Moura. Learning to understand image blur. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6586–6595, 2018. 1, 2, 3

[32] Z. Zhang, Y. Liu, Z. Xiong, J. Li, and M. Zhang. Focus and blurriness measure using reorganized dct coefficients for an autofocus application. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(1):15–30, 2018. 1

[33] W. Zhao, F. Zhao, D. Wang, and H. Lu. Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3080–3088, 2018. 1, 2, 3, 6, 7