



Hierarchical edge-aware network for defocus blur detection

Zijian Zhao^{1,2} · Hang Yang¹ · Huiyuan Luo¹

Received: 6 July 2021 / Accepted: 5 March 2022 / Published online: 29 March 2022
© The Author(s) 2022

Abstract

Defocus blur detection (DBD) aims to separate blurred and unblurred regions for a given image. Due to its potential and practical applications, this task has attracted much attention. Most of the existing DBD models have achieved competitive performance by aggregating multi-level features extracted from fully convolutional networks. However, they also suffer from several challenges, such as coarse object boundaries of the defocus blur regions, background clutter, and the detection of low contrast focal regions. In this paper, we develop a hierarchical edge-aware network to solve the above problems, to the best of our knowledge, it is the first trial to develop an end-to-end network with edge awareness for DBD. We design an edge feature extraction network to capture boundary information, a hierarchical interior perception network is used to generate local and global context information, which is helpful to detect the low contrast focal regions. Moreover, a hierarchical edge-aware fusion network is proposed to hierarchically fuse edge information and semantic features. Benefiting from the rich edge information, the fused features can generate more accurate boundaries. Finally, we propose a progressive feature refinement network to refine the output features. Experimental results on two widely used DBD datasets demonstrate that the proposed model outperforms the state-of-the-art approaches.

Keywords Defocus blur detection · Edge guidance aggregation · Hierarchical interior perception · Low contrast focal regions

Introduction

Defocus blur is a very common phenomenon in digital photos, arising from that the scene point is not at the camera's focal distance. Defocus blur detection (DBD) aims to distinguish blurred and unblurred regions from a given image. Defocus blur detection benefits much attention due to its practical applications such as salient object detection [1], defocus estimation [2], image restoration [3], blur region segmentation [4], and so on.

In the past decade, many defocus blur detection methods have been proposed. These methods can be simply divided into two categories: traditional methods and deep learning

based methods. The former one is based on hand-crafted features and utilizes low-level cues to predict DBD maps, such as frequency [5–9] and gradient [10–15]. However, these traditional methods can not well obtain global information of high-level semantic features; thus they can not accurately detect the low contrast focal regions (see green box region of Fig. 1a) and suppress the background clutter (see red box region of Fig. 1b). Otherwise, as shown in the blue box region of Fig. 1a, the boundaries of in-focus objects have not well been detected.

Recently, convolutional neural networks (CNNs) have been widely used in various computer vision tasks because of its powerful extraction capabilities, such as image denoising [19], image classification [20], super-resolution [21], salient object detection [22], and object tracking [23]. Similarly, CNNs have also been well applied in DBD [16,17,24–35]. Although deep learning based approaches achieve higher performance and significant improvements compared with the traditional methods, there remain several problems that need to be further addressed: (1) the complementary of local and global information generated by different layers can not be well utilized, which causes ambiguous detection of low-contrast regions and background clutter of the final DBD

✉ Hang Yang
yanghang@ciomp.ac.cn

Zijian Zhao
zhaozijian20@mails.ucas.ac.cn

Huiyuan Luo
luohuiyuan@ciomp.ac.cn

¹ Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science, Changchun 130033, China

² University of Chinese Academy of Sciences, Beijing 100049, China

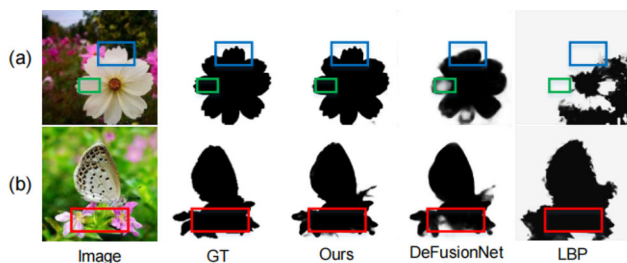


Fig. 1 Qualitative comparison of three models on the Shi's dataset [8] and DUT dataset [16], the first and the second columns show input images and their ground-truth images, respectively. From the third to the last columns, including Our DBD maps, DeFusionNet [17], and LBP [18]. Images in the green boxes are low contrast focal region patches, the red boxes are background clutter patches, and the blue boxes are the boundaries of in-focus objects patches

map; (2) the boundaries of in-focus objects can not be fully distinguished.

In this paper, we exploit a hierarchical edge-aware network (HEANet) to improve above-mentioned problems, which consists of four sub-networks: hierarchical interior perception network (HIPNet), edge feature extraction network (EFENet), hierarchical edge-aware fusion network (HEFNet), progressive feature refinement network (PFRNet). Specifically, considering the contextual information can benefit for detecting low contrast focal regions, we design a receptive field context module (RFCM) to capture multi-receptive field features. In addition, we cascade three RFCMs and form a top-bottom manner as the HIPNet. Then, we develop an EFENet to obtain the edge information of in-focus objects from feature maps. Subsequently, the multi-scale contextual features and the edge information are transmitted to the HEFNet, which consists of some progressive edge guidance aggregation modules (EGAMs). With this module, the edge cues and multi-scale semantic features can be hierarchically fused, making better performance on localization. Finally, we design a PFRNet to refine the feature maps to generate a DBD map with clear region boundaries, and supervise the predictive DBD map with the ground truth.

Our major contributions can be summarized as follows:

1. We propose a hierarchical edge-aware network (HEANet) for DBD, to the best of our knowledge, it is the first trial to develop an end-to-end network with edge awareness for DBD.
2. We design a receptive field context module (RFCM) to capture local and global context information, which aims to distinguish low contrast focal regions and suppress the background clutter. In addition, we cascade three RFCMs as HIPNet to extract the multi-scale contextual features hierarchically.
3. We develop an edge guidance aggregation module (EGAM), which incorporates edge information into the

hierarchical feature maps to guide the DBD maps to possess clear region boundaries.

4. Compared with 10 state-of-the-art approaches on two widely used datasets, our method outperforms the state-of-the-art approaches under five evaluation metrics.

Related work

In the past years, many DBD methods have been proposed. Traditional methods based on the hand-crafted features, such as frequency [5–9], gradient [10–15], and so on [18,36,37]. Shi et al. [8] propose a few local blur features, such as image gradient, Fourier domain, and data-driven local filters, to enhance the capabilities of defocus blur detection. Pang et al. [14] develop a new kernel-specific feature vector for DBD, which incorporates the multiplication of the variance of filtered kernel and the variance of filtered patch gradients. Yi et al. [18] present a sharpness metric based on local binary patterns to distinguish defocus regions. Tang et al. [36] design a blur metric based on the log averaged spectrum residual to obtain a coarse blur map, then an iterative updating mechanism is used to refine the blur map. Golestaneh et al. [37] propose a novel method based on high-frequency multi-scale fusion and sort transform of gradient magnitudes to compute blur detection maps. These traditional methods can be effective in some cases; however, they are the limited capacity to obtain high-level semantic information in complex scenarios.

Due to the powerful multi-level feature extraction capabilities, most deep learning based models can achieve better performance than traditional hand-crafted methods. In recent years, many approaches have adopted CNNs for DBD. Among these methods, Park et al. [25] propose a deep learning model to extract high-level features, then integrate the hand-crafted and high-level features to obtain a DBD map. Karaali et al. [24] develop an edge-based defocus blur estimation method. In this method, two CNNs are utilized to compute an edge map and estimate the unknown defocus blur amount, a fast image-guided filter is designed to propagate the sparse blur estimation to the whole image. However, each of these two methods is not a complete end-to-end network, the edges of the in-focus objects they generated are mostly blurry. Zhao et al. [16] adopt a multi-stream bottom-top-bottom fully convolutional network (BTBNet) to aggregate the multi-scale low-level and high-level features to predict the DBD map. Zhao et al. [26] propose a cross ensemble network to enhance the diversity of the features for DBD. Ma et al. [27] present an end-to-end local blur mapping algorithm for better detecting defocus blur regions. Lee et al. [28] develop a defocus map estimation network for spatially varying defocus map estimation and produced a novel depth-of-field dataset for the training network. Lately, Tang et al. [17] design a cross-layer structure to integrate low-level

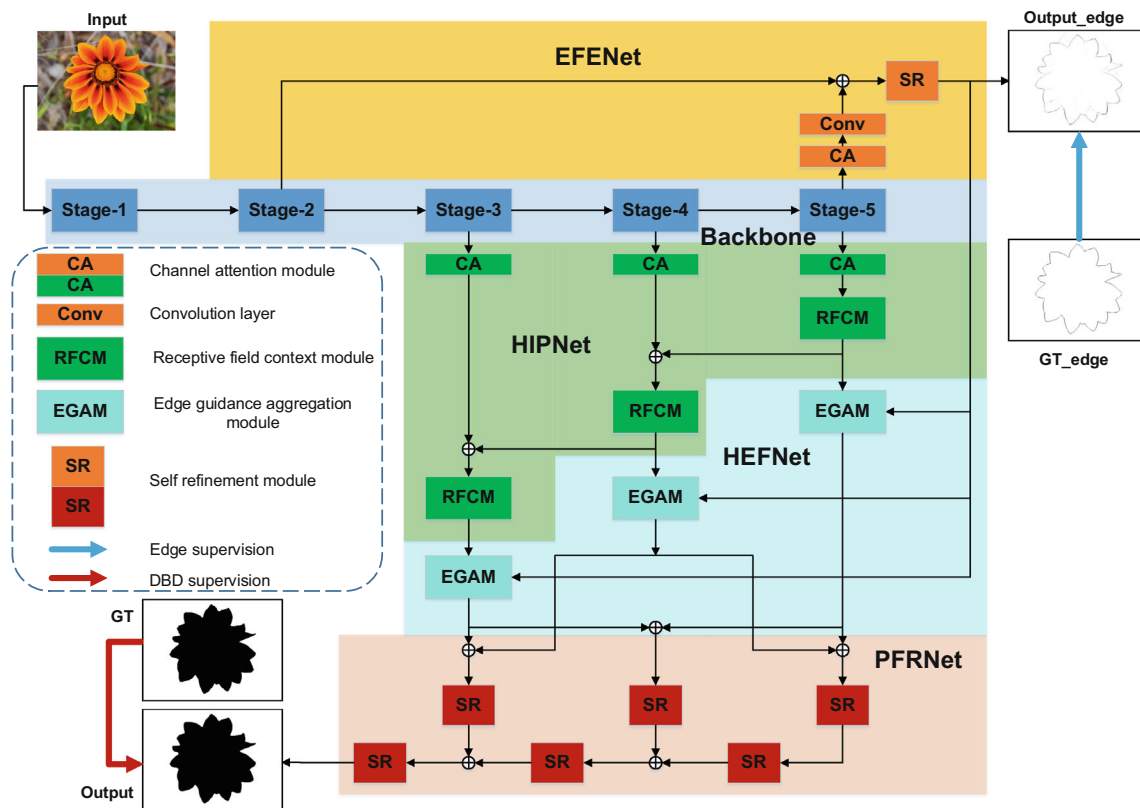


Fig. 2 The architecture of our HEANet. EFENet represents the edge feature extraction network. HIPNet is the hierarchical interior perception network. HEFNet represents the hierarchical edge-aware fusion network. PFRNet is the progressive feature refinement network

and high-level features step by step. Tang et al. [29] build a cross-layer framework and utilized an attention mechanism to integrate multi-level features. Tang et al. [30] propose a bidirectional residual feature refining method and introduce channel-wise attention to extract valuable features. Tang et al. [31] present a residual learning strategy to learn the residual maps, then use a recurrent method to combine the low-level and high-level features. Li et al. [32] design a complementary attention network by exploiting the complementary information of defocus feature maps. Zhao et al. [33] propose a cascaded DBD map residual learning architecture to recurrently refine the DBD maps. Zhao et al. [34] present two deep ensemble networks to boost diversity while costing less computation for DBD. Zhao et al. [35] adopt a method to train the model without using any pixel-level annotation that introduces dual adversarial discriminators, then, the generator is forced to generate an accurate DBD mask.

Inspired by but different from these approaches, in this paper, we concentrate on fusing the edge cues and semantic information hierarchically with a complementary mechanism. Experimental results show that our method has been achieved promising results.

Proposed HEANet

The framework of our method is illustrated in Fig. 2. Our approach includes four sub-networks: hierarchical interior perception network (HIPNet) which captures multi-scale contextual information, edge feature extraction network (EFENet) which extracts edge information, hierarchical edge-aware fusion network (HEFNet) which guides the extracted features hierarchical fusion by taking advantage of the edge information of low-level features, finally, progressive feature refinement network (PFRNet) is used to fuse and refine features progressively to generate the defocus blur map. These sub-networks consist of different modules. The details are introduced as follows.

Hierarchical interior perception network

The HIPNet consists of three channel attention modules (CA) [38] and three receptive field context modules (RFCM), first, we use CAs to reduce redundant information, then we cascade three RFCMs to hierarchically extract multi-scale contextual features from multi-level feature maps.

In HIPNet, the key requirement is to capture multi-scale contextual features. To expand such capability, we design

a receptive field context module (RFC-M) to extract multi-scale contextual information to detect the low contrast focal regions.

The proposed RFCM consists of 5 parallel branches, and we show the structure of RFCM in Fig. 3. First, we use 1×1 convolution to compress the channel of the feature map. Then, four branches from left to right, we employ a convolutional layer and dilated convolutional layer in each branch. The global convolutional network (GCN) [39] is utilized in the convolutional layer, we use GCNs with $k = 1, 3, 5, 7$ to obtain multi-scale features in the four branches. As shown in Fig. 4. The $k \times k$ convolutional operation is replaced by the combination of $k \times 1 + 1 \times k$ and $1 \times k + k \times 1$ convolutions to reduce parameters. In the dilated convolutional layer, we utilize 3×3 kernels but different dilation rates in the four branches to expend receptive fields and obtain local information. The dilation rates of the four dilated convolutional layers are set to $\{1, 3, 5, 7\}$, respectively. To obtain the successive dilation rates, we add three inter-branch short connections from the first branch to the fourth branch. In this way, the feature maps generated from the previous branch are encoded in the feature maps of subsequent branches. After that, the feature maps of four branches are up-sampled and concatenated, merging into a convolution array. Furthermore, an average pooling branch is adopted to obtain global information of feature maps. Finally, the convolution array of four branches and the output features of the pooling branch is integrated with an add operation, a ReLU layer is used to ensure the nonlinearity.

Edge feature extraction network

In this network, we intend to effectively extract edge features of in-focus objects. Inspired by the work of [40], we embed a channel attention (CA) module [38] to reduce the redundant information. The structure of CA is shown in Fig. 5. In order to enhance edge features, we embed self refinement (SR) module [41] on the side path to refine the final edge features. The structure of self refinement (SR) module is shown in Fig. 6. Specifically, the prediction of the edge map is supervised by the defocus blur edge ground truth.

Hierarchical edge-aware fusion network

We utilize EFENet to obtain low-level edge cues, and leverage three RFCMs in the HIPNet to hierarchically extract multi-scale contextual features at three different levels of backbone. These different levels have different discriminative information. High-level features have semantic and global information, these features can recognize the position of defocus blur regions. Low-level features retain spatial and local information, which can help divide the blur and clear regions.

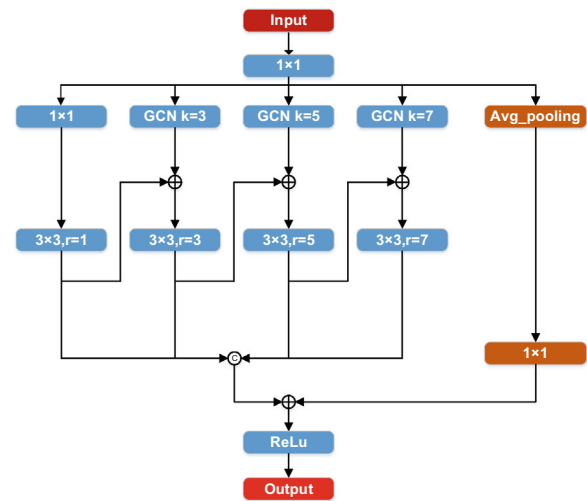


Fig. 3 The structure of receptive field context module (RFCM). “ $3 \times 3, r = 3$ ” represents the “ 3×3 ” convolutional operation and the dilation rate 3. “Avg_pooling” represents an average pooling operation. The symbol “c” denotes concatenation

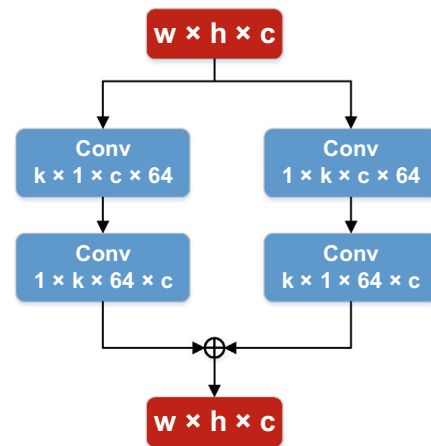


Fig. 4 The structure of global convolutional network (GCN). The $k \times k$ convolutional operation is replaced by the combination of $k \times 1 + 1 \times k$ and $1 \times k + k \times 1$ convolutions to reduce parameters



Fig. 5 The structure of channel attention (CA) module

After obtaining the low-level edge cues and high-level semantic features, we aim to leverage the edge information to guide the semantic features to perform better in localization. Therefore, as shown in Fig. 2, we develop an HEFNet, which uses multiple edge guidance aggregation modules (EGAMs) to embed the edge information into hierarchical feature maps, and guide them to possess clear region boundaries.

In order to integrate low-level edge cues and high-level semantic features effectively, we propose an edge guidance

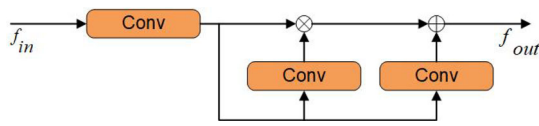


Fig. 6 The structure of self refinement (SR) module

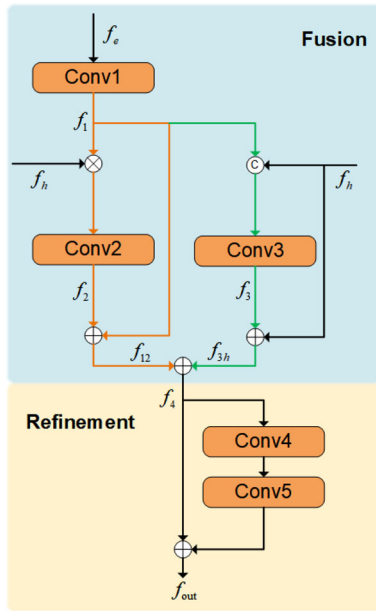


Fig. 7 The structure of edge guidance aggregation module (EGAM). f_e represents the input of edge features, f_h is the input of high level semantic features. f_{out} is the output of EGAM

aggregation module (EGAM). As shown in Fig. 7. EGAM receives two inputs, including the high-level features from the output of HIPNet, and the low-level edge cues from the EFENet. Specifically, its inner structure can be divided into two stages: fusion strategy and features refinement.

The fusion stage consists of two branches, from left to right, the first branch is to enhance the edge information of feature maps, we adopt the multiplication operation to strengthen the boundaries of defocus blur regions, meanwhile, suppressing the background noises. In this manner, we use the nature of edge cues f_e to guide semantic features f_h . At first, the channels of low-level edge features f_e are compressed to the same number of high-level features f_h through a 1×1 convolutional layer Conv1. Then, the edge cues f_1 and semantic features f_h through multiplication operation and feed into one 3×3 convolutional layer Conv2. Furthermore, the fused features f_2 will be added to the edge features f_1 for refine representations. The above process can be formulated as:

$$\begin{aligned}
 f_1 &= \text{Conv1}(f_e), & (1) \\
 f_2 &= \text{Conv2}(f_1 \times f_h), & (2) \\
 f_{12} &= f_2 + f_1. & (3)
 \end{aligned}$$

The second branch is to capture consistent semantics of high-level features. First, we combine edge features f_1 and high-level semantic features f_h by concatenation, one 3×3 convolutional layer Conv3 is used to obtain more local information, and then we add the fused feature f_3 to the high-level semantic features f_h . Further, the aggregated features f_{12} of the first branch and the features f_{3h} of the second branch will be added. The output of the fusion stage is then passed to the features refinement stage. The above process can be described as:

$$f_3 = \text{Conv3}(\text{concat}(f_1, f_h)), \tag{4}$$

$$f_{3h} = f_3 + f_h, \tag{5}$$

$$f_4 = f_{12} + f_{3h}. \tag{6}$$

As shown in Fig. 7, the features refinement stage also consists of two branches, one connects the input and output directly, the other branch consists of two 3×3 convolutional layers. Two branches are fused by an add operation, which is beneficial to learn the edge information and semantic information, thus the features f_4 from the first stage can be refined. The whole process can be defined as follows:

$$f_{out} = f_4 + \text{Conv5}(\text{Conv4}(f_4)). \tag{7}$$

With this design, the output of the first stage will obtain the properties of clear boundaries and consistent semantics. Each of the above-mentioned 3×3 convolutional layers consists of a convolutional layer with 3×3 kernel size, a batch normalization layer, and a ReLU layer. The output of the HEFNet is then fed to the PFRNet.

Progressive feature refinement network

In order to aggregate the multi-scale features from HEFNet effectively, we develop a PFRNet, which is inspired but different from coarse-to-fine residual learning in [33], the method in [33] only applies residual learning to reconstruct the output to the original resolution from the small scale to the large scale step by step. In our PFRNet, we combine coarse-to-fine residual learning and cross-level features fusion manner to enhance residual learning. At first, the multi-scale output features of EGAMs are cascaded fusion through cross-level features fusion manner as the input features of PFRNet, which are guiding the current step to learn residual features. Then, coarse-to-fine residual learning strategy is utilized to reconstruct the output to the original resolution through multiple SR modules. The SR module is used to refine and enhance the feature maps. After multiple adding operations and SR modules in PFRNet, we utilize a convolutional layer with 1×1 kernel size to obtain the final DBD map.

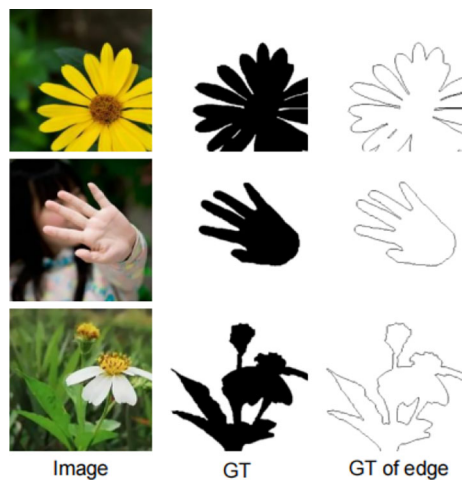


Fig. 8 Visualization of the ground truth of edge. The first and the second columns show input images and their ground-truth images, respectively. The last column shows the ground truth of edges, which are generated through the gradients of the ground truth of the images

Loss function

In defocus blur detection, binary cross-entropy (BCE) is widely used as a loss function, which calculates the loss between the final DBD map and ground truth. However, the BCE loss function does not consider the structural information of the defocus blur region, which may reduce the performance of the model. Inspired by the work of [42], we use a pixel position-aware (PPA) loss as our loss function, which is formed as:

$$L_{\text{ppa}}(p_{ij}, g_{ij}) = \alpha_{ij} \times L_{\text{bce}}(p_{ij}, g_{ij}) + L_{\text{wiou}}(p_{ij}, g_{ij}), \quad (8)$$

where p_{ij} and g_{ij} represent the DBD prediction and ground truth of the pixel (i, j) , respectively. L_{bce} is the binary cross-entropy loss, L_{wiou} is the weighted IOU loss. α_{ij} is the edge-aware weight, which is defined as:

$$\alpha_{ij} = 1 + \gamma \times |\text{avg_pool}(g_{ij}) - g_{ij}|, \quad (9)$$

where γ denotes the hyper-parameter, it is set as 5 in this work. L_{wiou} is formed as:

$$L_{\text{wiou}} = 1 - \frac{\alpha_{ij} \times \text{inter} + 1}{\alpha_{ij} \times \text{union} - \alpha_{ij} \times \text{inter} + 1}, \quad (10)$$

where $\text{inter} = p_{ij} \times g_{ij}$, and $\text{union} = p_{ij} + g_{ij}$.

The dominant loss of output corresponds to the L_{ppa} (p_{ij}, g_{ij}), we use the binary cross-entropy (BCE) loss as the edge loss function, the total loss is defined as:

$$L_{\text{total}} = L_{\text{ppa}}(p_{ij}, g_{ij}) + \lambda \times L_{\text{bce}}(pe_{ij}, ge_{ij}), \quad (11)$$

where λ represents the weight of different loss, λ is set to 0.3, $L_{\text{ppa}}(p_{ij}, g_{ij})$ and $L_{\text{bce}}(pe_{ij}, ge_{ij})$ denote the output loss and edge loss, respectively. The pe_{ij} and ge_{ij} are the edge prediction and ground truth of the edge pixel (i, j) , respectively.

Experiments

Datasets and evaluation metrics

Datasets The proposed approach is evaluated on two public blurred image datasets, including Shi [8], DUT [16]. Shi's dataset [8] is the earliest public blurred image dataset. There are 604 defocus blurred images for training and 100 defocus blurred images for testing. DUT [16] consists of 500 challenging defocus blurred images. There are complex background and low contrast focal regions in many images.

Evaluation metrics Five standard metrics are used to evaluate the model, including E-measure [43], S-measure [44], mean absolute error (MAE), precision and recall (PR) curve [8,24,37] and F-measure. E-measure metric is used to evaluate the similarity between the prediction and the ground truth. S-measure aims to evaluate region-aware and object-aware structural similarity between the defocus map and ground truth. More details about the E-measure and S-measure can be found in [43,44]. F-measure denotes an overall performance measurement, and it is formed as:

$$F_{\beta} = \frac{(1 + \beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}, \quad (12)$$

where β^2 is 0.3. MAE is used to evaluate the average difference between prediction map and ground truth, and it is defined as:

$$\text{MAE} = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W |P(i, j) - G(i, j)|, \quad (13)$$

W and H represent the width and height of images, respectively.

Implementation details

We utilize Pytorch to implement our model. ResNet-50 [45] is used as the backbone network, which is pre-trained on ImageNet. 604 defocus blurred images of Shi's dataset are used to train HEANet and other above-mentioned datasets are used to test HEANet. Our Method requires ground truth of regions and edges for training, while the above datasets can not provide the ground truth of edges. As shown in Fig. 8, the ground truth of edges is generated through the gradients of the ground

Table 1 Quantitative comparison including F-measure (F_β , larger is better), MAE (smaller is better), S-measure (larger is better) and E-measure (larger is better) over two widely used datasets

Method	Shi				DUT			
	F-measure	MAE	S-measure	E-measure	F-measure	MAE	S-measure	E-measure
DBDF [8]	0.841	0.324	0.851	0.581	0.803	0.364	0.468	0.543
SS [36]	0.835	0.266	0.602	0.553	0.866	0.246	0.611	0.552
KSFV [14]	0.733	0.380	0.427	0.311	0.746	0.400	0.439	0.332
LBP [18]	0.866	0.186	0.640	0.739	0.876	0.178	0.637	0.758
HiFST [37]	0.856	0.232	0.644	0.689	0.868	0.296	0.544	0.596
DMENet [28]	0.914	0.342	0.594	0.524	0.934	0.308	0.627	0.571
DBM [27]	0.917	0.155	0.734	0.772	0.779	0.283	0.459	0.424
DHDE [25]	0.850	0.390	0.544	0.463	0.822	0.405	0.508	0.442
BTBNet [16]	0.887	0.107	0.851	0.870	0.888	0.190	0.668	0.674
DeFusionNet [17]	0.914	0.117	0.757	0.845	0.923	0.119	0.732	0.803
DENets [34]	0.925	0.055	0.885	0.941	0.852	0.098	0.812	0.870
SG [35]	0.800	0.123	0.760	0.854	0.671	0.177	0.661	0.743
Ours	0.938	0.094	0.862	0.873	0.943	0.091	0.765	0.859

The best two results are marked in red, blue

truth of the images. For data augmentation, we use multi-scale, random crop, and horizontal flip input images. The initial learning rate is set to 0.05. We use stochastic gradient descent (SGD) to optimize the network. Warm-up and linear decay strategies are used to adjust the learning rate. Momentum and weight decay are set to 0.9 and 0.0005, respectively. The batch size is set to 10 and the whole training process is completed in 6K iterations with the maximum epoch of 101. The training process is about 1.5 h. Two RTX 3090 GPUs are used for acceleration. During testing, we resize each image to 320×320 and then feed it to HEANet to predict defocus blur maps without any post-processing.

Comparison with state-of-the-art methods

To evaluate the proposed HEANet, we compare it against 12 state-of-the-art algorithms, including defocus blur detection via recurrently fusing and refining multi-scale deep features (DeFusionNet) [17], defocus map estimation using domain adaptation (DMENet) [28], high-frequency multi-scale fusion and sort transform of gradient magnitudes (HiFST) [37], multi-scale deep and hand-crafted features for defocus estimation (DHDE) [25], local binary patterns (LBP) [18], discriminative blur detection features (DBDF) [8], spectral and spatial approach (SS) [36], multi-stream bottom–top–bottom fully convolutional network (BTBNet) [16], deep blur mapping via exploiting high-Level semantics (DBM) [27] and classifying discriminative features (KSFV) [14], defocus blur detection via boosting diversity of deep ensemble networks (DENets) [34], self-generated defocus blur detection via dual adversarial discriminators (SG) [35]. For the results of these methods except DENets and SG, we download the results from Tang’s [17] homepage. As for DENets and SG, we use the authors’ recommended and original implementations parameters.

Quantitative comparison Table 1 shows our method outperforms other approaches under four evaluation metrics, including F-measure, MAE, S-measure, and E-measure. Our model achieves the top two results on Shi’s dataset and DUT dataset for four metrics. It demonstrates the superior performance of our proposed HEANet. Fig. 9 shows the precision-recall curves of above-mentioned approaches on two datasets, from these curves, we can observe that the performance of HEANet is better than other models. It means that our method has a good capability to detect defocus blur regions as well as generate accurate defocus blur maps. *Qualitative comparison* In Fig. 10, we visualize some defocus blur maps produced by our model and other methods to evaluate the proposed HEANet. It can be seen that the HEANet clearly detects defocus blur regions and suppresses the background clutter. The HEANet is superior in handling a variety of challenging scenes, including low contrast focal regions (row 1 and row 6) and cluttered backgrounds (row 3 and row 4). Compared with other counterparts, the HEANet can not only distinguish the blur and clear regions but also retain their sharp boundaries. The edges of in-focus objects predicted by our HEANet are clearer, and the DBD maps are more accurate.

Ablation studies

The proposed HEANet contains four sub-networks, including the HIPNet, the EFENet, the HEFNet, and the PFRNet. Among them, the EFENet and the HEFNet are combined to extract and fuse edge information. In this section, we carry out a series of experiments to investigate the effectiveness of each component. The quantitative results of ablation studies are summarized in Table 2. In addition, the qualitative results are shown in Figs. 11 and 12.

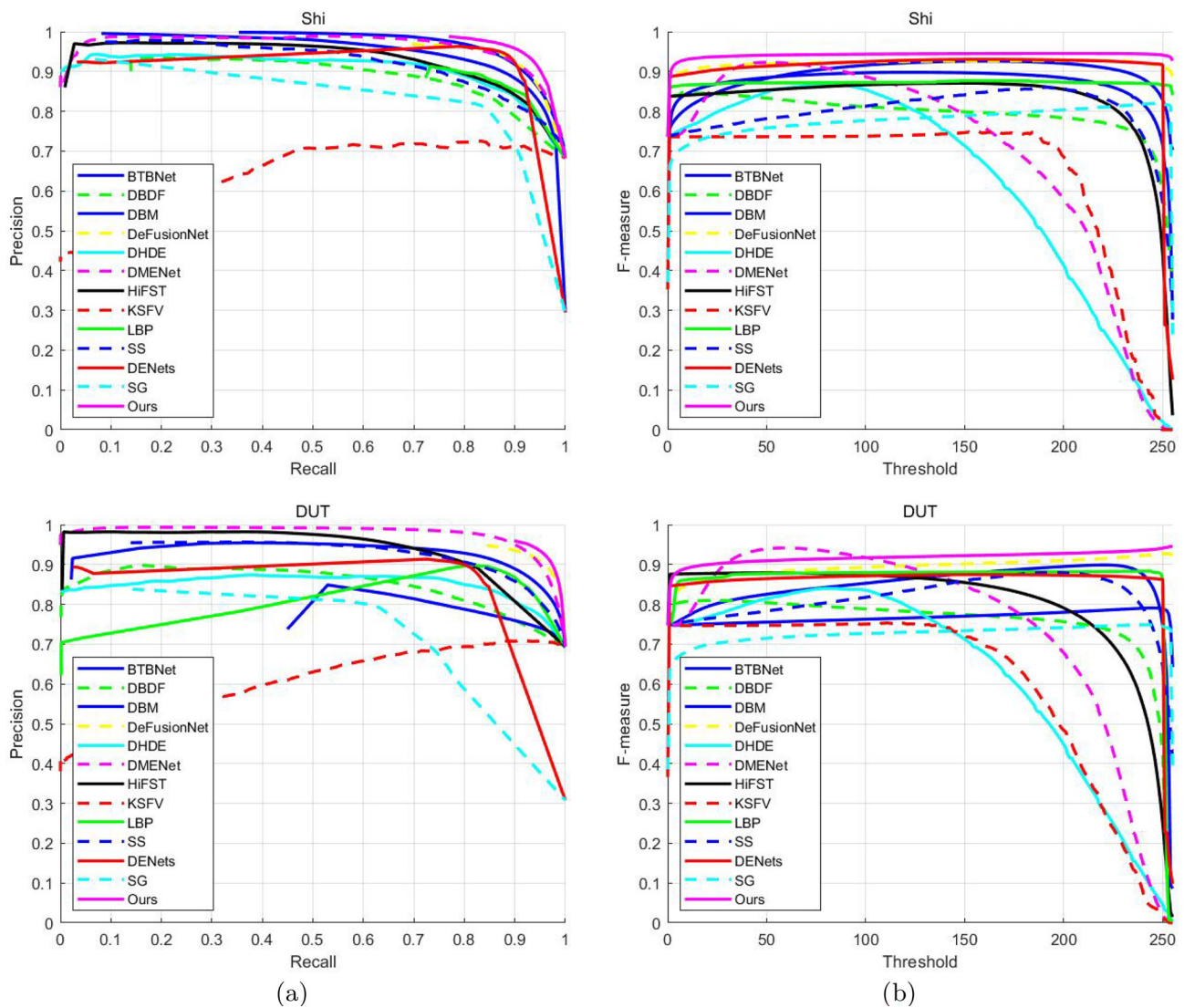


Fig. 9 PR and F-measure curves of 12 state-of-the-art methods over two datasets. The first row shows comparison of PR and F-measure curves on Shi's dataset [8]. The second row shows comparison of PR and F-measure curves on DUT dataset [16]

Effectiveness of HIPNet We utilize the HIPNet to capture multi-scale global contextual features, which is the key sub-network to detect low contrast focal regions. As it can be seen in the 1st and 2nd rows of Table 2, when we add the HIPNet to the backbone (HIPNet + ResNet-50), the quantitative results of HIPNet + ResNet-50 can comprehensively surpass the performances of ResNet-50. To further verify the effectiveness of the HIPNet, we show a visual comparison in Fig. 11. It can be seen that our proposed HIPNet is more deliberate to deal with the complex scene and can detect low contrast focal regions. Both results can illustrate the effect of the HIPNet in our model.

Effectiveness of EFENet and HEFNet The EFENet and HEFNet are the key sub-networks for our model to introduce and incorporate edge information, to investigate the

effect of our proposed EFENet and HEFNet, we have done two experiments across all two datasets comparisons. One is without EFENet and HEFNet, the other is embedded with EFENet and HEFNet. By comparing the 3rd and 5th rows of Table 2, the model embedded EFENet and HEFNet has much better performance than that without edge information. Several visual examples are shown in Fig. 12, with the help of EFENet and HEFNet, our method retains both accurate semantic information and edge information.

Effectiveness of PFRNet As shown in the 4th and 5th rows of Table 2, it can be observed that the model with PFRNet has a better performance than that without PFRNet.

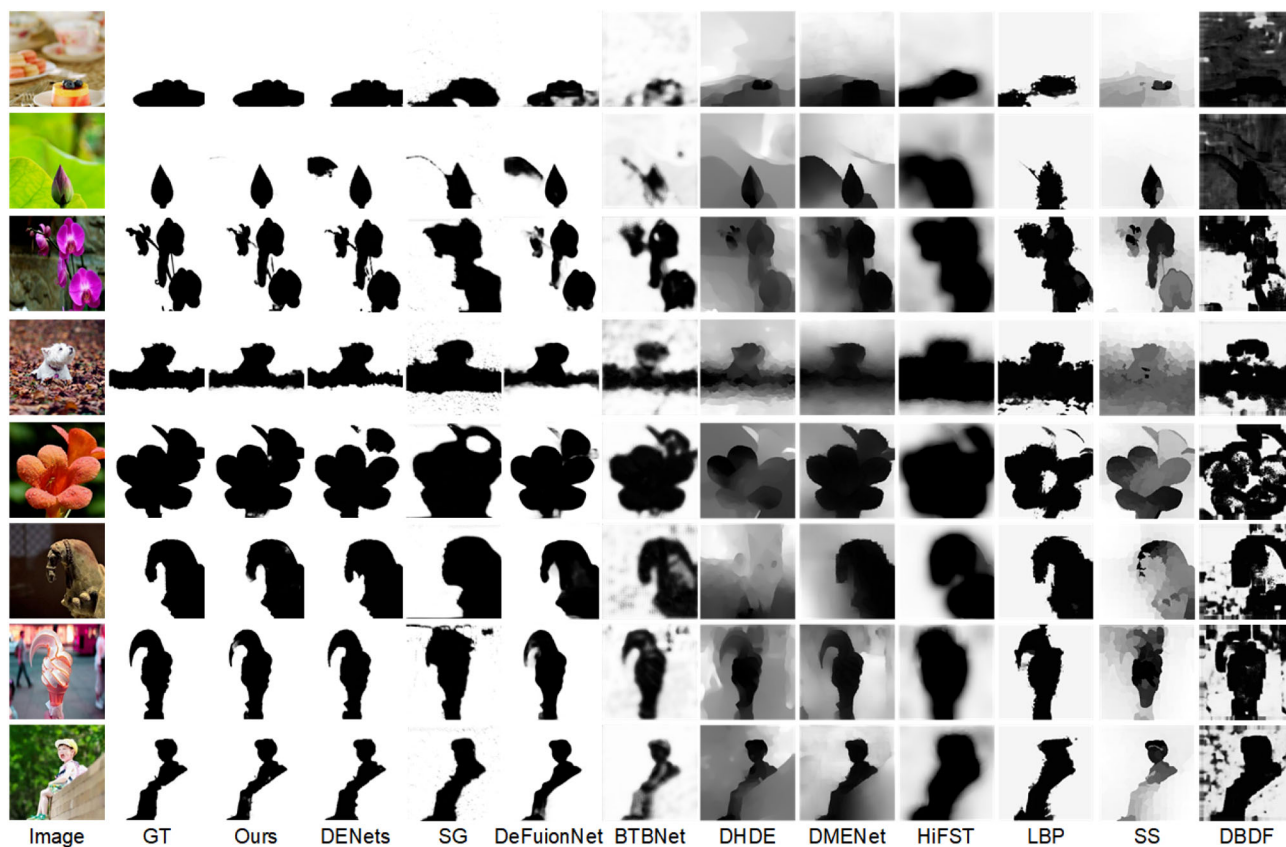


Fig. 10 Qualitative comparisons of the state-of-the-art methods and our approach. The first and the second columns show input images and their ground-truth images, respectively. The third column are the output images of our approach. The fourth to last columns are the state-of-the-art methods, including defocus blur detection via boosting diversity of deep ensemble networks (DENets) [34], self-generated defocus blur detection via dual adversarial discriminators (SG) [35], defocus blur detection via recurrently fusing and refining multi-scale

deep features (DeFusionNet) [17], multi-stream bottom–top–bottom fully convolutional network (BTBNet) [16], multi-scale deep and hand-crafted features for defocus estimation (DHDE) [25], defocus map estimation using domain adaptation (DMENet) [28], high-frequency multi-scale fusion and sort transform of gradient magnitudes (HiFST) [37], local binary patterns (LBP) [18], spectral and spatial approach (SS) [36] and discriminative blur detection features (DBDF) [8]

Table 2 Different module of ablation studies

ResNet-50	HIPNet	HEFNet	PFRNet	EFENet	Shi		DUT	
					F-measure	MAE	F-measure	MAE
✓					0.903	0.116	0.922	0.119
✓	✓				0.922	0.098	0.926	0.109
✓	✓		✓		0.927	0.096	0.927	0.105
✓	✓	✓		✓	0.928	0.095	0.929	0.103
✓	✓	✓	✓	✓	0.929	0.093	0.941	0.091

F-measure (F_β , larger is better), MAE (smaller is better). The best results are highlighted in red

Conclusion

In this paper, we propose a DBD approach named HEA-Net. To our knowledge, it is the first trial to develop an end-to-end network with edge awareness for defocus blur detection. First, we adopt an HIPNet to efficiently extract and aggregate multi-scale contextual information. Furthermore, the EFENet is used to capture the edge features of in-focus objects. Then, we propose an HEFNet to hierar-

chically fuse edge cues and semantic features to perform better in localization. Finally, we develop a PFRNet to refine the feature maps to generate a DBD map with clear edges. Experimental results demonstrate that our network outperforms state-of-the-art methods on two widely used datasets without any pre-processing or post-processing.

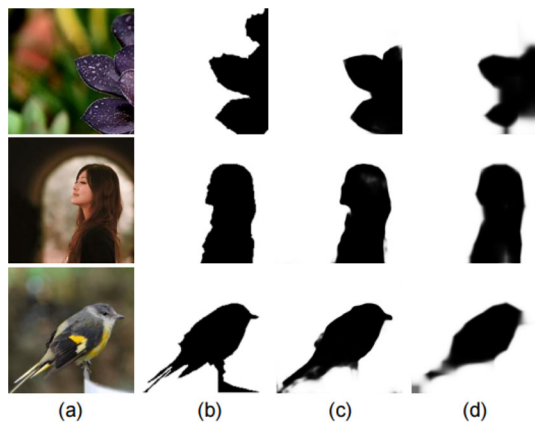


Fig. 11 Visual comparisons of our ablation studies. **a** input images, **b** ground truth, **c** results of backbone + HIPNet, **d** results of backbone

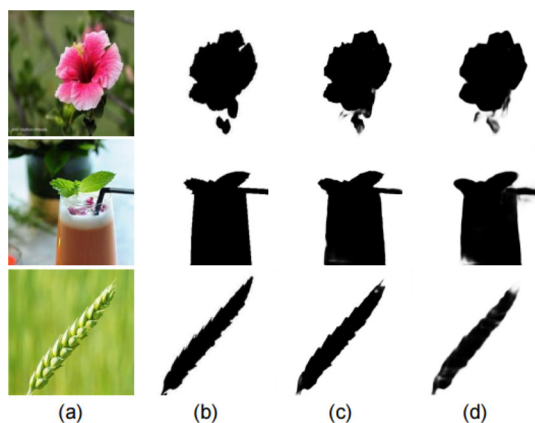


Fig. 12 Visual comparisons of our ablation studies. **a** input image, **b** ground truth, **c** with EFENet and HEFNet, **d** without EFENet and HEFNet

Acknowledgements This work is supported by the Chinese Academy of Sciences-Youth Innovation Promotion Association, Grant number 2020220, recipient Hang Yang; the National Natural Science Foundation of China (NSFC) Grant 62175086; and the Department of Science and Technology of Jilin Province (20210201132GX).

Declarations

Conflict of interest Corresponding authors declare on behalf of all authors that there is no conflict of interest. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the

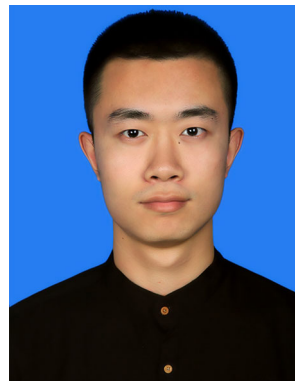
permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

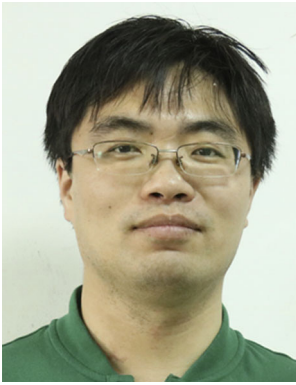
- Xia C, Gao X, Li KC et al (2020) Salient object detection based on distribution-edge guidance and iterative Bayesian optimization. *Appl Intell* 50:2977–2990. <https://doi.org/10.1007/s10489-020-01691-7>
- Tang C, Hou C, Song Z (2013) Defocus map estimation from a single image via spectrum contrast. *Opt Lett* 38(10):1706–1708. <https://doi.org/10.1364/OL.38.001706>
- Zhang X, Wang R, Jiang X et al (2016) Spatially variant defocus blur map estimation and deblurring from a single image. *J Vis Commun Image Represent* 35(1):257–264. <https://doi.org/10.1016/j.jvcir.2016.01.002>
- Levin A, Rav-Acha A, Lischinski D (2008) Spectral matting. *IEEE Trans Pattern Anal Mach Intell* 30(10):1699–1712. <https://doi.org/10.1109/TPAMI.2008.168>
- Zhu X, Cohen S, Schiller S et al (2013) Estimating spatially varying defocus blur from a single image. *IEEE Trans Image Process* 22(12):4879–4891. <https://doi.org/10.1109/TIP.2013.2279316>
- Vu CT, Phan TD, Chandler DM (2012) S_3 : a spectral and spatial measure of local perceived sharpness in natural images. *IEEE Trans Image Process*. <https://doi.org/10.1109/TIP.2011.2169974>
- Zhang Y, Hirakawa K (2013) Blur processing using double discrete wavelet transform. *IEEE Conf Comput Vis Pattern Recognit*. <https://doi.org/10.1109/CVPR.2013.145>
- Shi J, Xu L, Jia J (2014) Discriminative blur detection features. *IEEE Conf Comput Vis Pattern Recognit*. <https://doi.org/10.1109/CVPR.2014.379>
- Tang C, Wu J, Hou Y, Wang P, Li W (2016) A spectral and spatial approach of coarse-to-fine blurred image region detection. *IEEE Signal Process Lett*. <https://doi.org/10.1109/LSP.2016.2611608>
- Zhuo S, Sim T (2011) Defocus map estimation from a single image. *Pattern Recognit* 44(9):1852–1858. <https://doi.org/10.1016/j.patcog.2011.03.009>
- Zhao J, Feng H, Xu Z et al (2013) Automatic blur region segmentation approach using image matting. *Signal Image Video Process* 7(6):1173–1181. <https://doi.org/10.1007/s11760-012-0381-6>
- Su B, Lu S, Tan CL (2011) Blurred image region detection and classification. *ACM international conference on multimedia*, pp 1397–1400
- Saad E, Hirakawa K (2016) Defocus blur-invariant scale-space feature extractions. *IEEE Trans Image Process* 25(7):3141–3156. <https://doi.org/10.1109/TIP.2016.2555702>
- Pang Y, Zhu H, Li X et al (2017) Classifying discriminative features for blur detection. *IEEE Trans Cybern* 46(10):2220–2227. <https://doi.org/10.1109/TCYB.2015.2472478>
- Liu R, Li Z, Jia J (2008) Image partial blur detection and classification. *IEEE Conf Comput Vis Pattern Recognit*. <https://doi.org/10.1109/CVPR.2008.4587465>
- Zhao W, Zhao F, Wang D et al (2018) Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network. *IEEE conference on computer vision and pattern recognition*, pp 3080–3088. <https://doi.org/10.1109/CVPR.2018.00325>
- Tang C, Zhu X, Liu X et al (2019) DeFusionNET: defocus blur detection via recurrently fusing and refining multi-scale deep features. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 2695–2704. <https://doi.org/10.1109/CVPR.2019.00281>

18. Xin Y, Eramian M (2016) LBP-based segmentation of defocus blur. *IEEE Trans Image Process* 25(4):1–1. <https://doi.org/10.1109/TIP.2016.2528042>
19. Zhang K, Zuo W, Chen Y et al (2017) Beyond a gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Trans Image Process.* <https://doi.org/10.1109/TIP.2017.2662206>
20. Wei Y, Wei X, Min L et al (2016) HCP: a flexible CNN framework for multi-label image classification. *IEEE Trans Softw Eng* 38(9):1901–1907. <https://doi.org/10.1109/TPAMI.2015.2491929>
21. Dong C, Loy CC, He K et al (2016) Image super-resolution using deep convolutional networks. *IEEE Trans Pattern Anal Mach Intell* 38(2):295–307. <https://doi.org/10.1109/TPAMI.2015.2439281>
22. Jiao J, Xue H, Ding J (2021) Non-local duplicate pooling network for salient object detection. *Appl Intell.* <https://doi.org/10.1007/s10489-020-02147-8>
23. Li P, Wang D, Wang L et al (2018) Deep visual tracking: review and experimental comparison. *Pattern Recognit* 76:323–338
24. Karaali A, Harte N, Jung CR (2020) Deep multi-scale feature learning for defocus blur estimation. [arXiv:2009.11939](https://arxiv.org/abs/2009.11939)
25. Park J, Tai Y W, Cho D et al (2017) A unified approach of multi-scale deep and hand-crafted features for defocus estimation. *IEEE Computer Society*, pp 2760–2769
26. Zhao W, Zheng B, Lin Q et al (2019) Enhancing diversity of defocus blur detectors via cross-ensemble network. *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp 8897–8905. <https://doi.org/10.1109/CVPR.2019.00911>
27. Ma K, Fu H, Liu T et al (2016) Deep blur mapping: exploiting high-level semantics by deep neural networks. *IEEE Trans Image Process.* <https://doi.org/10.1109/TIP.2018.2847421>
28. Lee J, Lee S, Cho S et al (2019) Deep defocus map estimation using domain adaptation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 12214–12222. <https://doi.org/10.1109/CVPR.2019.01250>
29. Tang C, Liu X, Zheng X et al (2020) DeFusionNET: defocus blur detection via recurrently fusing and refining discriminative multi-scale deep features. *IEEE Trans Pattern Anal Mach Intell.* <https://doi.org/10.1109/TPAMI.2020.3014629>
30. Tang C, Liu X, An S et al (2021) BR²Net: defocus blur detection via a bidirectional channel attention residual refining network. *IEEE Trans Multimed.* <https://doi.org/10.1109/TMM.2020.2985541>
31. Tang C, Liu X, Zhu X et al (2020) R²MRF: defocus blur detection via recurrently refining multi-scale residual features. *Proc AAAI Conf Artif Intell* 34(7):12063–12070. <https://doi.org/10.1609/aaai.v34i07.6884>
32. Li J, Fan D, Yang L et al (2021) Layer-output guided complementary attention learning for image defocus blur detection. *IEEE Trans Image Process.* <https://doi.org/10.1109/TIP.2021.3065171>
33. Zhao W, Zhao F, Wang D et al (2020) Defocus blur detection via multi-stream bottom-top-bottom network. *IEEE Trans Pattern Anal Mach Intell.* <https://doi.org/10.1109/TPAMI.2019.2906588>
34. Zhao W, Hou X, He Y et al (2021) Defocus blur detection via boosting diversity of deep ensemble networks. *IEEE Trans Image Process.* <https://doi.org/10.1109/TIP.2021.3084101>
35. Zhao W, Shang C, Lu H (2021) Self-generated defocus blur detection via dual adversarial discriminators. *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp 6929–6938. <https://doi.org/10.1109/CVPR46437.2021.00686>
36. Tang C, Wu J, Hou Y et al (2016) A spectral and spatial approach of coarse-to-fine blurred image region detection. *IEEE Signal Process Lett* 23(11):1652–1656. <https://doi.org/10.1109/LSP.2016.2611608>
37. Golestaneh SA, Karam LJ (2017) Spatially-varying blur detection based on multiscale fused and sorted transform coefficients of gradient magnitudes. *IEEE conference on computer vision and pattern recognition*, pp 5800–5809. <https://doi.org/10.1109/CVPR.2017.71>
38. Hu J, Shen L, Albanie S et al (2017) Squeeze-and-excitation networks. *IEEE Trans Pattern Anal Mach Intell.* <https://doi.org/10.1109/TPAMI.2019.2913372>
39. Peng C, Zhang X, Yu G, et al. (2017) Large kernel matters-improve semantic segmentation by global convolutional network. *IEEE conference on computer vision and pattern recognition (CVPR)*, pp 1743–1751. <https://doi.org/10.1109/CVPR.2017.189>
40. Zhao J, Liu J, Fan D et al (2020) EGNet: edge guidance network for salient object detection. *IEEE/CVF international conference on computer vision (ICCV)*, pp 8778–8787. <https://doi.org/10.1109/ICCV.2019.00887>
41. Chen Z, Xu Q, Cong R et al (2020) Global context-aware progressive aggregation network for salient object detection. *Proc AAAI Conf Artif Intell* 34(7):10599–10606. <https://doi.org/10.1609/aaai.v34i07.6633>
42. Wei J, Wang S, Huang Q (2019) F3Net: fusion, feedback and focus for salient object detection. [arXiv:1911.11445](https://arxiv.org/abs/1911.11445)
43. Fan D, Gong C, Yang C et al (2018) Enhanced-alignment measure for binary foreground map evaluation, pp 698–704. <https://doi.org/10.24963/ijcai.2018/97>
44. Fan D, Cheng M, Liu Y et al (2017) Structure-measure: a new way to evaluate foreground maps. *IEEE international conference on computer vision (ICCV)*, pp 4558–4567. <https://doi.org/10.1109/ICCV.2017.487>
45. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. *Proc IEEE Conf Comput Vis Pattern Recognit.* <https://doi.org/10.1109/CVPR.2016.90>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Zijian Zhao received the B.S. degree from Changchun University of Technology in 2017. He is currently studying toward his M.S. degree at Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science. His research interest includes visual detection.



Hang Yang received his B.S. and Ph.D. degrees in mathematics from the Jilin University in 2007 and 2012, respectively. He is currently an Associate Researcher at the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science. His current research interests include image deblurring and visual tracking.



Huiyuan Luo received the B.S. degree from Harbin Institute of Technology, Weihai in 2016. He received his Ph.D. degree at Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science in 2021. His current research interests are mainly focused on saliency detection and deep learning.