Applied Intelligence (2022) 52:14426–14439
https://doi.org/10.1007/s10489-022-03303-y

# Defocus Blur detection via transformer encoder and edge guidance

**Zijian Zhao[1,2] · Hang Yang[1] · Huiyuan Luo[1]**

Accepted: 25 January 2022 / Published online: 8 March 2022


**Abstract**

Defocus blur detection (DBD) aims to separate blurred and unblurred regions for a given image. Benefiting from the powerful extraction capabilities of convolutional neural networks (CNNs), deep learning based defocus blur detection has achieved a remarkable progress compared with traditional methods. However, due to the limited local receptive field of CNNs, it is difficult to achieve satisfactory results in the detection of the low-contrast focal regions. Besides, the output maps of the most of previous works have coarse object boundaries and background clutter. In this paper, we propose a hybrid CNN-Transformer architecture with an edge guidance aggregation module (EGAM) and a feature fusion module (FFM) for DBD. In our knowledge, this is the first study to utilize a transformer encoder for DBD to capture the global context information. Additionally, an edge extraction network (EENet) is adopted to obtain local edge information of in-focus objects. To effectively aggregate local edge information and global semantic features, three EGAMs are integrated into an edge guidance fusion network (EGFNet). Benefiting from the rich edge information, the fused features can generate more accurate boundaries. Finally, three FFMs are cascaded as a hierarchical feature aggregation network (HFANet) to hierarchically decode and refine the feature maps. Experimental results on three widely used DBD datasets demonstrate that the proposed model outperforms the state-of-the-art approaches.

**Keywords** Defocus blur detection · Edge guidance aggregation · Transformer encoder · Low-contrast focal regions

## 1 Introduction

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

In the past years, many defocus blur detection methods have been proposed. These methods can be simply divided into two categories: traditional methods and deep learning based methods. Hand-crafted features are adopted to predict DBD maps in early works, such as frequency features [5–10] and gradient features [11–16]. However, these methods can not well obtain high-level semantic features, thus they can not accurately detect the low-contrast focal regions (see green box region in Fig. 1(a)) and suppress the background clutter (see blue box region in Fig. 1(b)). Otherwise, as shown in the red box region in Fig. 1(a), the boundaries of in-focus objects are not clearly detected. Deep learning based approaches are dominated by CNNs, which have been widely used in various computer vision tasks, such as salient object detection [17], image denoising [18], super resolution [19], object tracking [20] and image classification [21]. Similarly, CNNs have also been successfully applied in DBD [22–31]. Although these deep learning approaches achieve higher performance compared with the traditional methods, there remains two problems that need to be further addressed: (1) the global context information can not be well obtained, which causes ambiguous detection of low-contrast regions and background clutter of the final DBD map; (2) the boundaries of in-focus objects can not be fully distinguished.

Recently, the transformer networks have achieved significant progress in many computer vision tasks, such as

✉ Hang Yang
yanghang@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

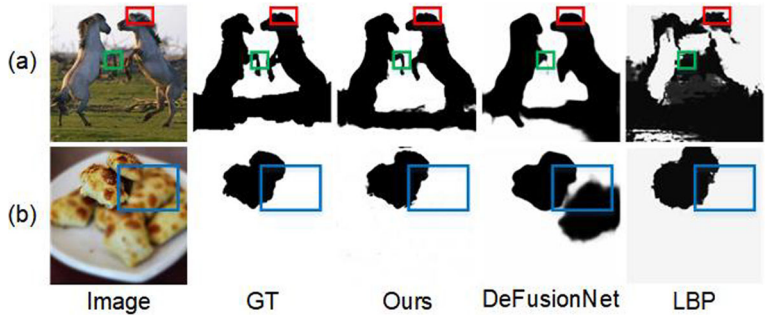



**Fig. 1** Several problems of defocus blur detection. From left to right: input image, ground truth (GT), our DBD maps, DeFusionNet [27], and LBP [39]

salient object detection [35, 48], image classification [43], image segmentation [33, 44], object tracking [49], etc. The advantage of the transformer encoder is the self-attention mechanism, which captures global context information at all stages to model a long-range dependency. Thus we introduce a transformer encoder to capture global context information of a given image, which helps the detection of low-contrast regions and suppresses background clutter.

In this paper, we focus on the detection of low-contrast regions and the distinguishment of in-focus objects boundaries. To remedy above mentioned problems, we propose a method based on transformer encoder and edge guidance, which consists of four components: hybrid CNN-Transformer backbone, edge extraction network (EENet), edge guidance fusion network (EGFNet), and hierarchical feature aggregation network (HFANet). The core of the first problem is the complementarity between global contextual information and localized spatial information. Specifically, we use CNNs to model local detailed features to perform better in localization. However, each convolution operation only focuses on the local area of the image, the features obtained through the convolution layer are not globally sensitive. Different from the limited local receptive field of CNNs, the transformer encoder has powerful modeling capabilities of long-range dependency to obtain global contextual information. Considering the powerful modeling capabilities of the long-range dependency of the transformer encoder, we introduce the transformer encoder to capture long-range dependency, which token image patches from CNNs as the input to obtain the global context of features. Thus a hybrid CNN-Transformer architecture is adopted from a top-bottom manner as our backbone. Then, we develop an EENet to capture the edge information of in-focus objects from feature maps. Subsequently, the contextual features and the edge information are transmitted to the EGFNet, which consists of several progressive edge guidance aggregation modules (EGAMs). With this module, the edge cues and semantic features can be effectively fused. In addition, we design a feature fusion module (FFM) to aggregate and refine the feature maps. Finally, different from the original decoders of the U-shape structure, we cascade three FFMs and form a bottom to top manner as the HFANet to generate a DBD map with clear regions boundaries and supervise the predictive DBD map with the ground-truth.

In summary, we propose a hybrid CNN-Transformer architecture from a top-bottom manner as our backbone. In our knowledge, this is the first attempt to use a transformer encoder for DBD to capture the global context information, which helps to detect low-contrast regions and suppress the background clutter. Besides, it is also the first trail to incorporate edge information into the feature maps to guide the DBD maps to possess clear regions boundaries. Compared with 10 state-of-the-art approaches on three datasets, our method outperforms other approaches with five evaluation metrics.

## 2 Related works

**Hand-crafted features based DBD** In early work- s, many DBD methods mostly use hand-crafted features to predict DBD maps, such as gradient based methods [11–16], frequency based methods [5–10], and so on [39–41]. These methods can be effective in some cases, however, it has limited capacity to obtain high-level semantic information in complex scenarios.

**Deep learning based DBD** In recent years, deep learning based models can achieve better performance than traditional hand-crafted approaches in DBD. Am- ong these methods, Park et al. [22] use a deep learning model to extract high-level features, then integrate the hand-crafted and high-level features to obtain a DBD map. However, this method is not a complete end-to-end network, the edges of the in-focus objects they generated are mostly blurry. In [23], a multi-stream bottom-top-bottom fully convolutional network (BTBNet) is proposed, which aggregates the multi-scale low-level and high-level features, and gradually refines the preceding blur detection maps to obtain a final DBD map. Ensemble networks [24] and high-level semantics [25] are also proposed for DBD. In [26], a novel depth-of-field dataset is produced for the training network. Besides, Tang et al. [28] present a cross-layer manner to

integrate low-level and high-level features to predict DBD maps. In [27], a cross-layer structure is proposed to progressively fuse and refine shallow features and deep features, and a channel attention module is designed to select discriminative features. Tang et al. [29] propose a bidirectional residual feature refining method and introduce channel-wise attention to extract valuable features. Tang et al. [30] propose 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. [31] propose a complementary attention network, which exploits the complementary information among each defocus image for DBD.

**Transformers in vision** Currently, the transformer networks have achieved remarkable progress in many computer vision tasks. For example, Vision Transformer (ViT) [43] is the first attempt to use a standard Transformer directly in computer vision and achieves great success on ImageNet classification. In [46], the ViT is utilized as an encoder to encode features and the convolutional layers are used as a decoder to progressively upsample and fuse features to obtain the final dense prediction. In [45], the progressive shrinking pyramid and spatial-reduction attention are proposed to capture high-resolution feature maps. In [47], it is the first study to use transformers for medical image segmentation, a hybrid CNN-Transformer architecture is deployed to encode detailed spatial features and global contextual information, the U-Net decoder is used to recover localized spatial information. [44] utilizes a transformer encoder to model an image as a sequence of patches, several traditional stacked convolution layers are deployed to recover the original image resolution.

In this paper, we concentrate on two aspects: the extraction of localized spatial features and global context information, fusing the edge cues and semantic information hierarchically with a complementary mechanism. Experiment shows that our method has been achieved promising results.

## 3 Proposed method

The framework of our method is illustrated in Fig. 2. Our approach is based on transformer encoder and edge guidance, which includes four components: hybrid CNN-Transformer backbone which captures the global context information and localized spatial features, edge extraction network (EENet) which obtains local edge information of in-focus objects, edge guidance fusion network (EGFNet) which guides the extracted features hierarchical fusion by taking advantage of the edge information. Finally, a hierarchical feature aggregation network (HFANet) is used to decode and fuse features hierarchically to generate the defocus blur map. The details are described as follows.

### 3.1 Hybrid CNN-transformer backbone

Our backbone concludes two parts: ResNet-50 [42] and transformer encoder [43]. ResNet-50 is firstly applied to downsample and extract different level local features of input image. Then, the transformer encoder takes image patches from ResNet-50 as the input to obtain global context of features.

**Transformer encoder** Inspired by the work of [43], which is the first attempt to generate 2D images by a basic transformer architecture. In this framework, the input image resolution H × W × 3 is reshaped into a sequence of flattened 2D patches $x_p$, each image patches of size is P × P, $N = \frac{HW}{P^2}$ is the number of image patches and each patch is flatted into a vector of size $\frac{3HW}{P^2}$. The transformer encoder is adopted to process global contextual information in our work, which is helpful to distinguish the low-contrast regions. As shown in Fig. 3. The structure of transformer encoder is composed of L transformer layers. Each transformer layer consists of multi-head self-attention layers (MSA) and multi-layer perceptron (MLP) blocks. More details of Transformer encoder can be found in [43]. The main process of $l$-th layer can be introduced as:

$$z_0 = [x_p^1 E; x_p^2 E; x_p^3 E; ...; x_p^N E] + E_{pos} \tag{1}$$

$$z_l' = MSA(LN(z_{l-1})) + z_{l-1} \tag{2}$$

$$z_l = MLP(LN(z_l')) + z_l' \tag{3}$$

where $E$ is the patch embedding projection, $E_{pos}$ is the position embedding, and $LN(\cdot)$ denotes layer normalization.

### 3.2 Edge extraction network

In this network, we intend to effectively extract edge features of in-focus objects. Different from the work of [34], we embed a channel attention (CA) module [32] to reduce the redundant information. The structure of CA is shown in Fig. 4. In order to enhance edge features, we embed the feature fusion module (FFM) on the side path to refine the final edge features. Specifically, the prediction of the edge map is supervised by the defocus blur edge ground-truth.

The proposed FFM is shown in Fig. 5. In detail, the features $f_u$ and $f_d$ are concatenated and fed into a 3 × 3 convolutional layer Conv1 to obtain features $f_1$. Then, one 3 × 3 convolutional layer Conv2 is applied to features $f_1$ to obtain features $f_2$. In addition, a residual structure is used to embed previous each convolutional layer of the FFM to keep the details of feature maps. In addition, the multiplication operation is adopted to strengthen the response of features and suppress the background noises. Further, the mix features of $f_1$ and $f_2$ are fed into Conv3

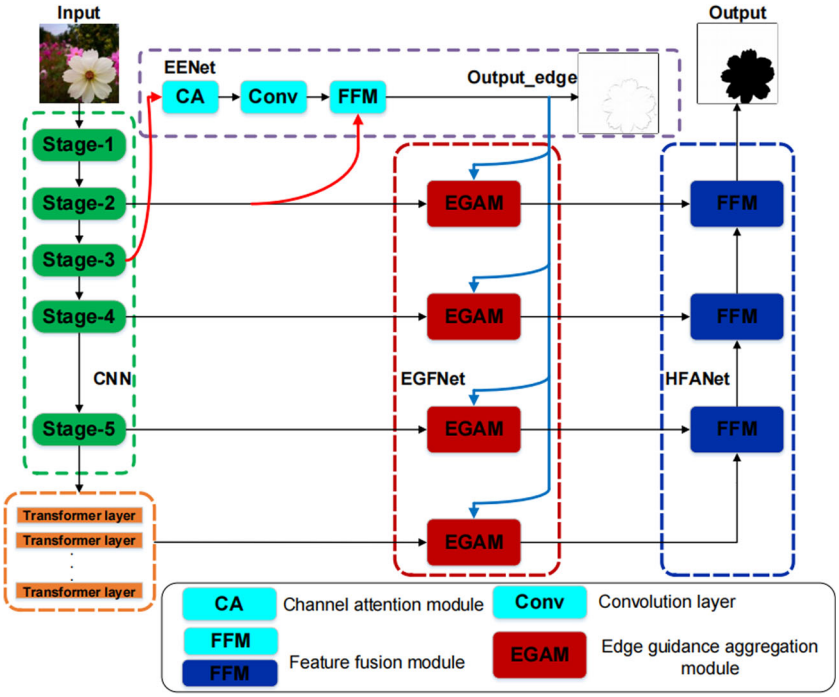



**Fig. 2** The architecture of our method. EENet represents the edge extraction network. EGFNet is the edge guidance fusion network. HFANet represents the hierarchical feature aggregation network

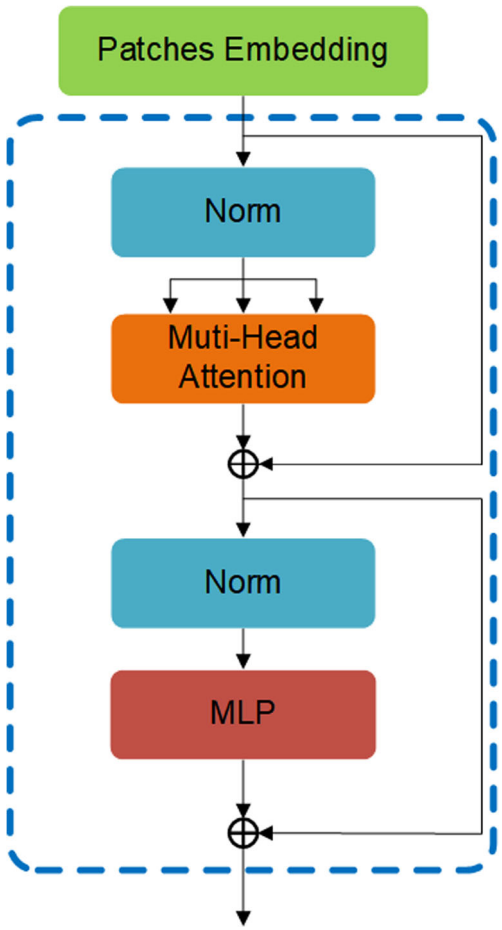



**Fig. 3** The structure of transformer layer

to obtain features $f_3$. Finally, an add operation is used to produce features $f_{out}$ as the output features of FFM. The whole process is introduced as :

$$f_1 = Conv1(concat(f_u, upsample(f_d))) \tag{4}$$

$$f_2 = Conv2(f_1) \tag{5}$$

$$f_3 = Conv3(f_1 \times f_2) \tag{6}$$

$$f_{out} = f_2 + f_3 \tag{7}$$

where "$concat$" denotes concatenation, "$upsample$" is bilinear interpolation. $Conv1$, $Conv2$, $Conv3$ are equipped with $3 \times 3$ kernel size, a batch normalization layer and a ReLU layer.

### 3.3 Edge guidance fusion network

We utilize the CNN-Transformer backbone to obtain local detailed features through different levels of CNN and global context information through transformer encoder. These different levels have different discriminative information. Global contextual features have semantic information, these features can recognize the position of defocus blur regions. Local features retain detailed spatial information, which can help divide the blur and clear regions.

After obtaining the edge cues and semantic features, we aim to utilize the edge features to guide the semantic features to perform better in localization. Therefore, as shown in Fig. 2, we develop an EGFNet, which uses multiple edge guidance aggregation modules (EGAMs) to

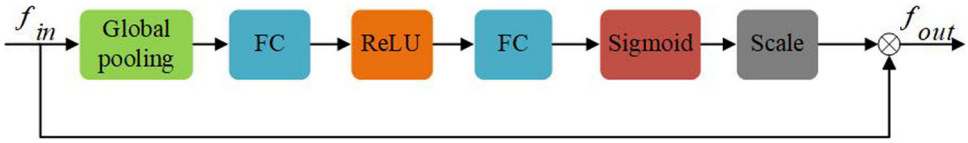



**Fig. 4** The structure of channel attention (CA) module

embed the edge information into hierarchical feature maps, and guide them to possess clear regions boundaries.

In order to integrate edge cues and high-level semantic features effectively, we propose an edge guidance aggregation module (EGAM). As shown in Fig. 6. The EGAM receives two inputs, including the semantic features $f_h$ from the output of the hybrid backbone, and the edge cues $f_e$ from the output of EENet. Specifically, its inner structure can be divided into two stages: features fusion and features refinement.

In the features fusion stage, we use the nature of edge features $f_e$ to guide semantic features $f_h$. Firstly, a $1 \times 1$ convolutional layer $Conv1$ is used for edge features $f_e$ to obtain the same channels as semantic features $f_h$. Then, edge cues $f_1$ and semantic features $f_h$ through multiplication operation and feed into one $3 \times 3$ convoluional layer $Conv2$. The multiplication operation is utilized to strengthen the boundaries of defocus blur regions, meanwhile suppress the background noises. Further, we add the middle features $f_2$ and the edge features $f_1$ to enhance the edge information of feature maps. The above process can be described as:

$$f_1 = Conv1(f_e) \tag{8}$$

$$f_2 = Conv2(f_1 \times f_h) \tag{9}$$

$$f_{12} = f_2 + f_1 \tag{10}$$

Besides, a mirror method is utilized to alleviate the effect of semantic features dilution. We combine $f_1$ and $f_h$ by concatenation, one $3 \times 3$ convolutional layer $Conv3$ is used to obtain more local information. Then, we add $f_h$ and middle features $f_3$. In addition, the aggregated features $f_{12}$ and $f_{3h}$ will be fused by add operation. The above process can be formulated as:

$$f_3 = Conv3(concat(f_1, f_h)) \tag{11}$$

$$f_{3h} = f_3 + f_h \tag{12}$$

$$f_4 = f_{12} + f_{3h} \tag{13}$$

Further, the output features $f_4$ is then passed to the features refinement stage. In the features refinement stage, it consists of two branches, one connects the input and output directly, the other branch consists of two $3 \times 3$ convolution layers $Conv4$, $Conv5$. Two branches are fused by an add operation, which is beneficial to learn the edge information and semantic information, thus the features from the features fusion stage can be refined. The features refinement process can be defined as follows:

$$f_{out} = f_4 + Conv5(Conv4(f_4)) \tag{14}$$

where "$concat$" denotes concatenation, $Conv1$ is $1 \times 1$ kernel size. $Conv2$, $Conv3$, $Conv4$, $Conv5$ are equipped with $3 \times 3$ kernel size, a batch normalization layer and a ReLU layer.

With this design, the features of EGAM will obtain the properties of clear boundaries and consistent semantics. The output of the EGFNet is then fed to the HFANet.

### 3.4 Hierarchical feature aggregation network

In order to aggregate the guided multi-scale features from EGFNet effectively, we develop an HFANet, in this network, we embed the FFM as the features refinement module. This module is used to refine and enhance the feature maps. After multiple FFMs in progressive feature refinement network, we utilize a convolutional layer with $1 \times 1$ kernel size to obtain the final DBD map.

### 3.5 Loss function

In DBD, the binary cross-entropy (BCE) is widely used as 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. In this paper, we introduce a pixel position-aware (PPA) loss [36] as our loss function, which is formed as:

$$L_{ppa}(p_{ij}, g_{ij}) = \alpha_{ij} \times L_{bce}(p_{ij}, g_{ij}) + L_{wiou}(p_{ij}, g_{ij}) \tag{15}$$

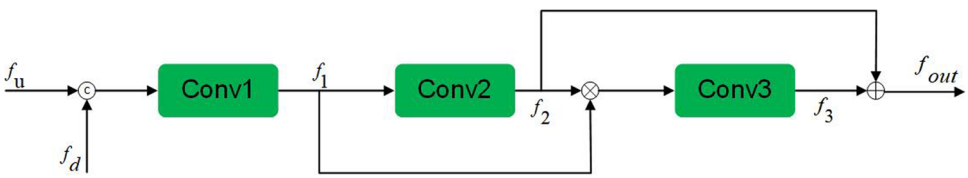



**Fig. 5** The structure of feature fusion module (FFM). The symbol "c" denotes concatenation. $f_u$ and $f_d$ represent two different input feature maps

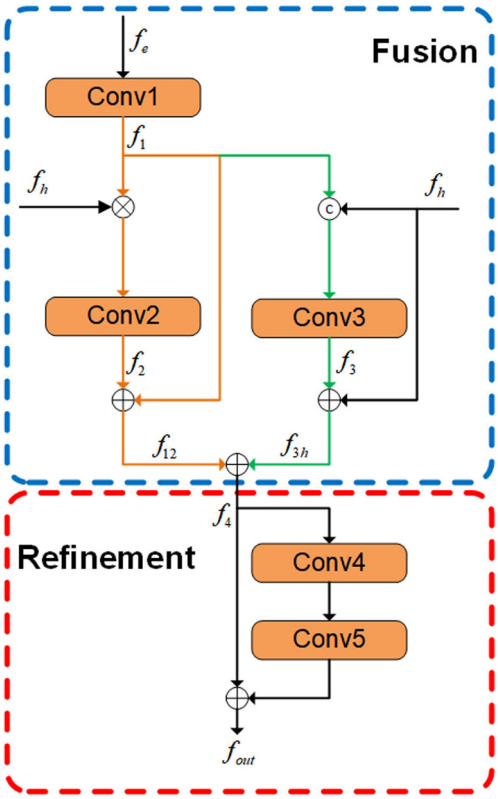



**Fig. 6** The structure of edge guidance aggregation module (EGAM). $f_e$ represents the input of edge features, $f_h$ is the input of high-level semantic features. $f_{out}$ is the output of EGAM

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. $\alpha_{ij}$ is the edge-ware weight, which is defined as :

$$\alpha_{ij} = 1 + \gamma \times |avg\_pool(g_{ij}) - g_{ij}| \tag{16}$$

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

$$L_{wiou} = 1 - \frac{\alpha_{ij} \times inter + 1}{\alpha_{ij} \times union - \alpha_{ij} \times inter + 1} \tag{17}$$

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

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:

$$L_{total} = L_{ppa}(p_{ij}, g_{ij}) + \lambda \times L_{bce}(pe_{ij}, ge_{ij}) \tag{18}$$

where $\lambda$ represents the weight of different loss, $\lambda$ is set to 0.3, $L_{ppa}(p_{ij}, g_{ij})$ and $L_{bce}(pe_{ij}, ge_{ij})$ 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.

# 4 Experiments

## 4.1 Datasets and evaluation metric

**Datasets** The proposed method is evaluated on three public blurred image datasets, including Shi [8], DUT [23]. 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 [23] consists of 500 challenging defocus blurred images. There are complex background and low-contrast focal regions in many images. CTCUG [27] is a new defocus blur detection dataset which contains 150 images with manual pixel-wise annotations. There are in-focus background with blurry foreground and complex background in many images.

**Evaluation metric** Five standard metrics are used to evaluate the model, including E-measure [38], S-mea- sure [37], mean absolute error (MAE), precision and recall (PR) curve [8, 10, 41] and F-measure. F-measure denotes an overall performance measurement, and it is formed as:

$$F = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall} \tag{19}$$

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

$$MAE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} |P(i, j) - G(i, j)| \tag{20}$$

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

## 4.2 Implementation details

We utilize Pytorch to implement our model. A hybrid CNN-Transformer architecture is used as the backbone network, ResNet-50 [42] and transformer encoder [43] with 12 transformer layers are pre-trained on ImageNet. 604 defocus blurred images of Shi's dataset are used to train the model and other above-mentioned datasets are used to test the model. The input images are resized as 320 × 320 and the patch size is set as 16. Our method requires ground truth of regions and edges for training, while the above datasets can not provide the ground truth of edges. The ground truth of edges is generated through the gradients of the ground truth of the images. The ground truth of edges is shown in Fig. 7. For data augmentation, we use random crop and horizontal flip input images. The initial learning rate is set to 0.01. We use the 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. Batch size is set to

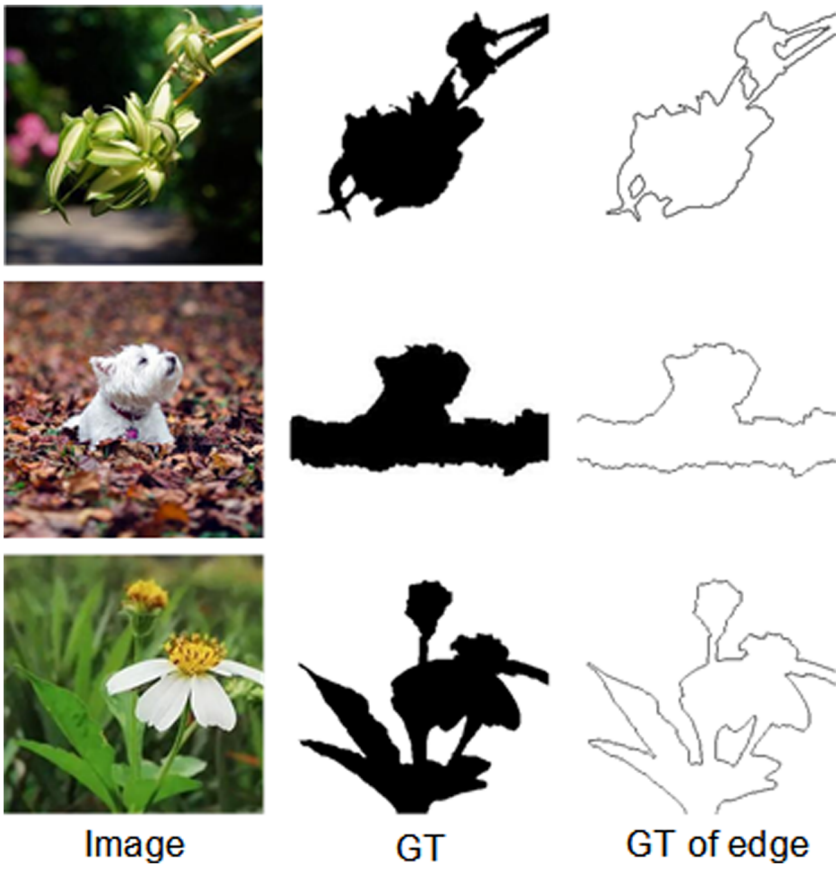



**Fig. 7** Visualization of the ground truth of edge

10. Two RTX 3090 GPUs are used for acceleration. During testing, we resize each image to 320 × 320 and the patch is set as 16, then feed it to our model to predict defocus blur maps without any post-processing.

### 4.3 Comparison with state-of-the-art methods

In order to evaluate the proposed method, we compare it against 10 state-of-the-art algorithms, including Defocus Blur Detection via Recurrently Fusing and Refining Multi-scale Deep Features(DeFusionNet) [27], defocus map estimation using domain adaptation (DMENet) [26], high-frequency multi-scale fusion and sort transform of gradient magnitudes (HiFST) [41], multi-scale deep and hand-crafted features for defocus estimation (DHDE) [22], local binary patterns (LBP) [39], discriminative blur detection features (DBDF) [8], spectral and spatial approach (SS) [40], multi-stream bottom-top-bottom fully convolutional network (BTBNet) [23], deep blur mapping via exploiting high-Level semantics (DBM) [25] and classifying discriminative features (KSFV) [15]. For the results of these methods, we download them from Tang's [27] homepage, which uses the authors' recommended and original implementations parameters.

**Quantitative comparison** Tables 1 and 2 show our method outperforms other approaches under four evaluation metrics, including F-measure, MAE, S-meas- ure and E-measure. Our model achieves the best scores on Shi and DUT datasets with respect to four metrics, compared with other counterparts, achieves the top two results on the CTCUG dataset. It demonstrates the superior performance of the proposed approach. Figure 8 shows the precision-recall curves of above-mentioned approaches on three datasets. From these curves, we can observe that the performance of our model is better than other approaches. 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. 9, we visualize so- me defocus blur maps produced by our model and other methods to evaluate the proposed model. It can be seen that our method clearly detects defocus blur regions and suppresses the background clutter. The proposed model is superior in handling a variety of challenging scenes, including low-contrast focal regions (row 3 and row 8) and cluttered backgrounds (row 5 and row 6). Compared with other counterparts, our method can not only distinguish the blur and clear regions, but also retain their sharp boundaries. The edges of in-focus

**Table 1** Quantitative comparison including F-measure (larger is better), MAE (smaller is better), S-measure (larger is better) and E-measure (larger is better) over Shi and DUT datasets

| Method | Shi | | | | DUT | | | |
|---|---|---|---|---|---|---|---|---|
| | F-meature | MAE | S-meature | E-meature | F-meature | MAE | S-meature | E-meature |
| DBDF | 0.841 | 0.324 | 0.851 | 0.581 | 0.803 | 0.364 | 0.468 | 0.543 |
| SS | 0.835 | 0.266 | 0.602 | 0.553 | 0.866 | 0.246 | 0.611 | 0.552 |
| KSFV | 0.733 | 0.380 | 0.427 | 0.311 | 0.746 | 0.400 | 0.439 | 0.332 |
| LBP | 0.866 | 0.186 | 0.640 | 0.739 | 0.876 | 0.178 | 0.637 | 0.758 |
| HiFST | 0.856 | 0.232 | 0.644 | 0.689 | 0.868 | 0.296 | 0.544 | 0.596 |
| DMENet | 0.914 | 0.342 | 0.594 | 0.524 | **0.934** | 0.308 | 0.627 | 0.571 |
| DBM | **0.917** | 0.155 | 0.734 | 0.772 | 0.779 | 0.283 | 0.459 | 0.424 |
| DHDE | 0.850 | 0.390 | 0.544 | 0.463 | 0.822 | 0.405 | 0.508 | 0.442 |
| BTBNet | 0.887 | **0.107** | **0.851** | **0.870** | 0.888 | 0.190 | 0.668 | 0.674 |
| DeFusionNet | 0.914 | 0.117 | 0.757 | 0.845 | 0.923 | **0.119** | **0.732** | **0.803** |
| **Ours** | **0.943** | **0.087** | **0.788** | **0.888** | **0.947** | **0.069** | **0.796** | **0.859** |

The best two results are marked in **red** and **blue**

objects predicted by our method are clearer, and the DBD maps are more accurate.

## 4.4 Ablation studies

Our proposed method consists of four sub-networks namely the hybrid backbone, the EENet, the EGFNet, and the HFANet. Among them, the EENet and the EGFNet are combined to extract and fuse edge information. In this section, we carry out a serial of experiments to investigate the effectiveness of each component of the model. The quantitative results of ablation studies are summarized in Table 3. In addition, the qualitative results are shown in Figs. 10 and 11. Furthermore, we add the multi-scale supervision (MSS) to evaluate the effect of supervision, the quantitative results of different supervision are summarized in Table 4.

**Effectiveness of transformer encoder** We utilize the transformer encoder to capture contextual features, which helps the model expand the receptive field and get global information to detect low-contrast focal regions. We use a visual

**Table 2** Quantitative comparison including F-measure (larger is better), MAE (smaller is better), S-measure (larger is better) and E-measure (larger is better) over a new dataset CTCUG

| | CTCUG | | | |
|---|---|---|---|---|
| Method | F-meature | MAE | S-meature | E-meature |
| DBDF | 0.740 | 0.345 | 0.496 | 0.654 |
| SS | 0.796 | 0.288 | 0.591 | 0.695 |
| KSFV | 0.607 | 0.461 | 0.391 | 0.393 |
| LBP | 0.805 | 0.243 | 0.600 | 0.728 |
| HiFST | 0.785 | 0.267 | 0.592 | 0.702 |
| DMENet | 0.846 | 0.301 | 0.639 | 0.794 |
| DBM | 0.832 | 0.209 | 0.658 | 0.799 |
| DHDE | 0.811 | 0.307 | 0.612 | 0.761 |
| BTBNet | 0.827 | 0.177 | 0.675 | 0.769 |
| DeFusionNet | **0.852** | **0.131** | **0.725** | **0.814** |
| Ours | **0.850** | **0.124** | **0.730** | **0.813** |

The two best results are marked in **red** and **blue**

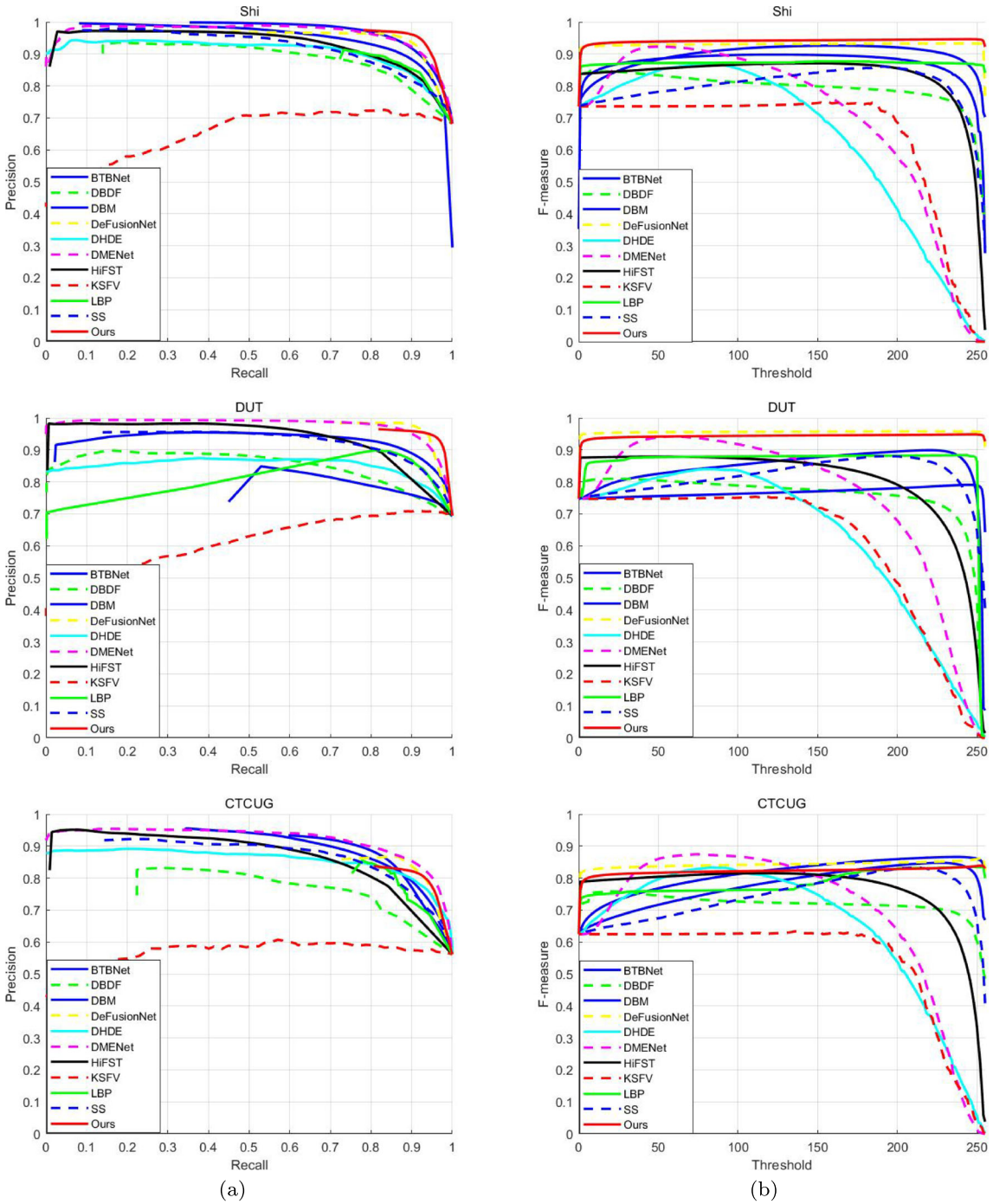



**Fig. 8** PR and F-measure curves of 10 state-of-the-art methods over three datasets. The first column shows comparison of PR curves. The second column shows comparison of F-measure curves of different methods on three datasets

comparison to verify the effectiveness of the transformer encoder, as shown in the 3rd and 4th columns of Fig. 10. It can be seen, when we add the transformer encoder to the backbone, in-focus and defocus blur regions will be more distinct and background clutter will be suppressed. Furthermore, as seen in the 1st and 3rd rows of Table 3, it has a beneficial effect on DBD and improves the results.

**Effectiveness of EENet and EGFNet** To investigate the effectiveness of our proposed EENet and EGF- Net, we

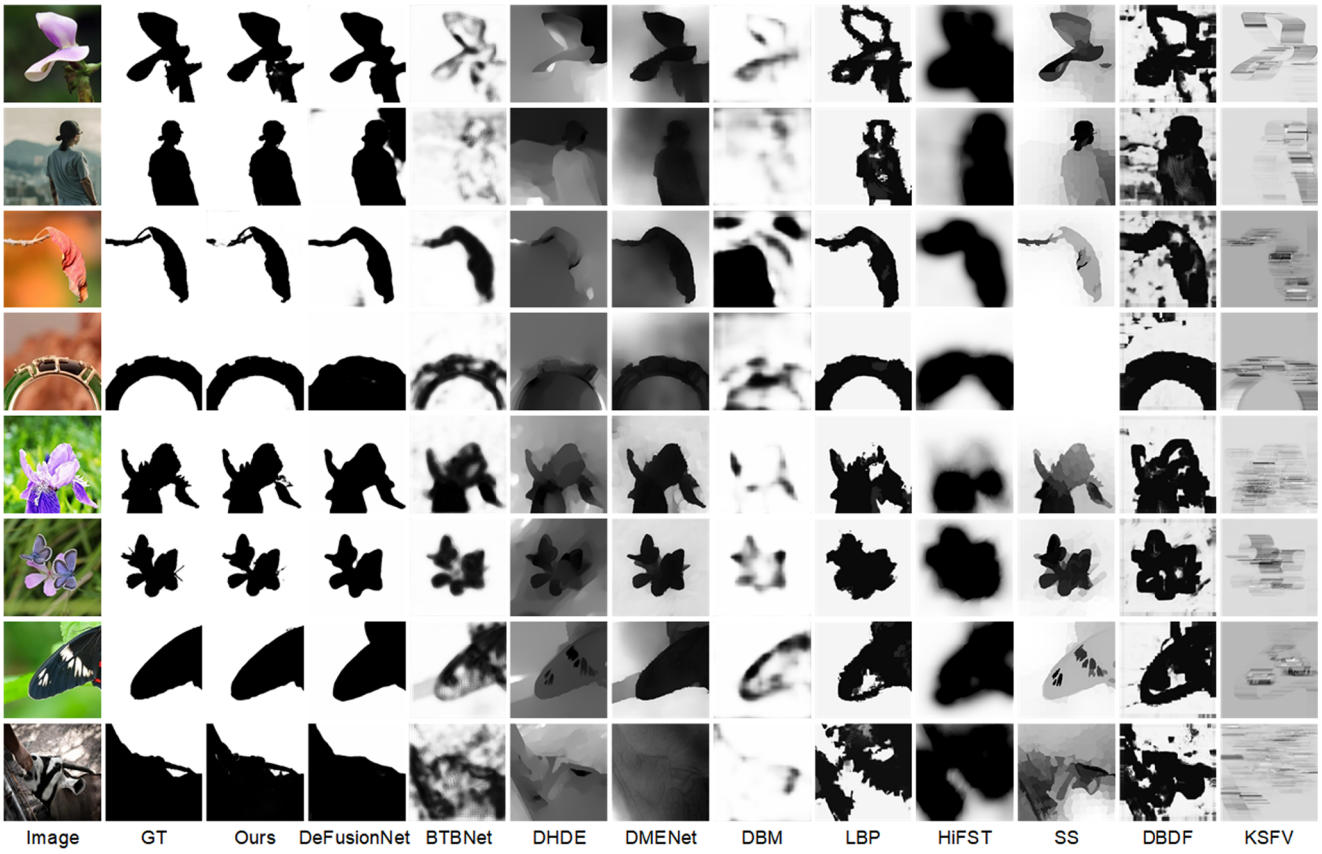



**Fig. 9** Qualitative comparisons of the state-of-the-art methods and our approach

conduct ablation experiments across all three datasets by introducing two different settings for comparisons. One is without edge information, the other is embedded with EENet and EGFNet. By comparing the 4th and 6th rows of Table 3, the model embedded EENet and EGFNet has much better performance. The EENet detects and refines the boundaries of in-focus objects, the EGFNet uses the nature of edge features to guide semantic features and fuses them hierarchically. Several visual examples are illustrated in the 4th and 6th columns of Fig. 11. With the help of EENet and EGFNet, our method retains both accurate semantic information and edge information.

**Effectiveness of HFANet** As shown in the 5th and 6th rows of Table 3, it can be observed that the model with HFANet has a better performance than that without HFANet.

**Table 3** Different module of ablation studies

| ResNet-50 | Hybrid Backbone | EENet | EGFNet | HFANet | Shi | | DUT | |
|---|---|---|---|---|---|---|---|---|
| | | | | | F-measure | MAE | F-measure | MAE |
| √ | | | | | 0.924 | 0.116 | 0.927 | 0.119 |
| √ | | √ | √ | √ | 0.937 | 0.094 | 0.932 | 0.107 |
| | √ | | | | 0.935 | 0.100 | 0.945 | 0.075 |
| | √ | | | √ | 0.939 | 0.092 | 0.946 | 0.072 |
| | √ | √ | √ | | 0.942 | 0.089 | 0.946 | 0.071 |
| | √ | √ | √ | √ | **0.943** | **0.087** | **0.947** | **0.069** |

The best results are highlighted in **red**

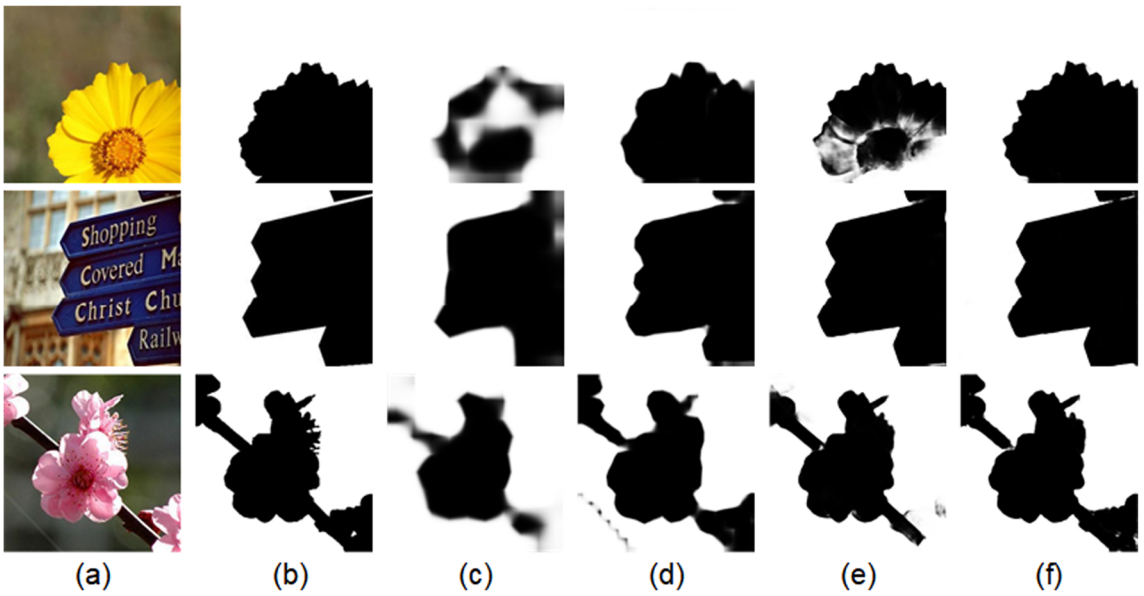



**Fig. 10** Visual comparisons of our ablation studies. (a) input image, (b) ground truth, (c) results of ResNet50, (d) results of hybrid backbone, (e) ResNet50 + EENet + EGFNet + HFANet, (f) hybrid backbone + EENet + EGFNet + HFANet

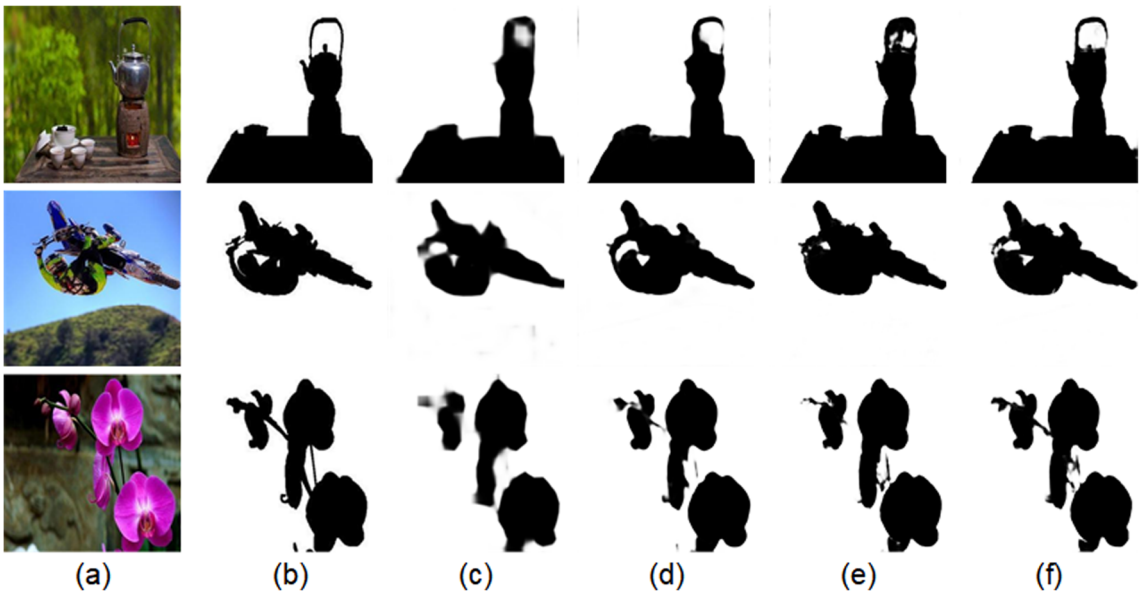



**Fig. 11** Visual comparisons of our ablation studies. (a) input image, (b) ground truth, (c) results of hybrid backbone, (d) results of hybrid backbone + HFANet, (e) hybrid backbone + EENet + EGFNet, (f) hybrid backbone + EENet + EGFNet + HFANet

**Table 4** Different supervision of ablation studies. MSS means multi-scale supervision

| MSS | Hybrid Backbone | EENet | EGFNet | HFANet | Shi | | DUT | | CTCUG | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | F-measure | MAE | F-measure | MAE | F-measure | MAE |
| | √ | √ | √ | √ | 0.943 | 0.087 | 0.947 | 0.069 | **0.850** | **0.124** |
| √ | √ | √ | √ | √ | **0.948** | **0.081** | **0.952** | **0.061** | 0.813 | 0.147 |

The best results are highlighted in **red**

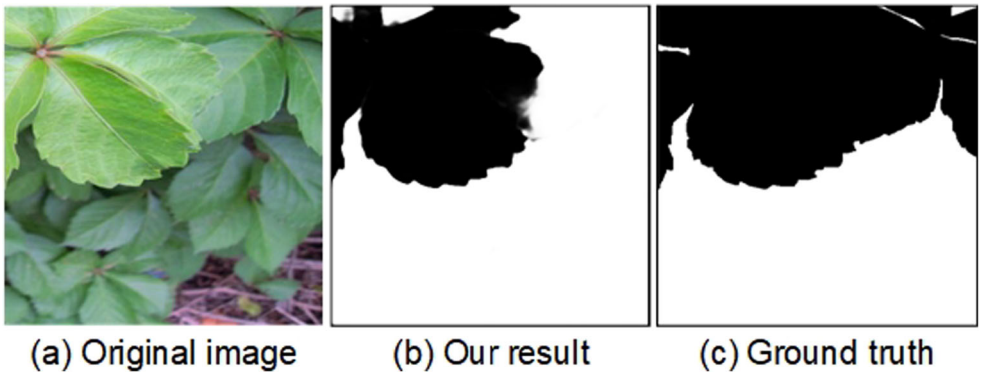



**Fig. 12** One failure example

**Effectiveness of supervision** The MSS loss consists of two parts, i.e., the FFM loss of HFANet and the loss of edge information. Among them, the PPA loss function is utilized as FFM loss, the binary cross-entropy (BCE) loss is used as the edge loss function. The MMS loss is defined as:

$$L_{mms} = \sum_{i=1}^{3} \lambda_i L^i_{ffm} + \mu L_{edge} \tag{21}$$

where $L_{edge}$ means the edge loss, $L^i_{ffm}$ denotes the loss of the output of the i-*th* FFM in HFANet, $\mu$ and $\lambda_i$ represent the weight of different loss, $\mu$ is set to 0.3, inspired by the work of [36], the weight of $\{\lambda_1, \lambda_2, \lambda_3\}$ corresponds to {1, 1/4, 1/8}, respectively.

As shown in Table 4, it can be observed that the model with MSS has better results than that with single supervision in the Shi dataset and DUT dataset. However, the model with MSS has worse performance in the CTCUG dataset. This is because the data distribution of the CTCUG dataset is different from the Shi dataset and DUT dataset. In the Shi dataset and DUT dataset, most of the images have in-focus foreground regions and blurry backgrounds. In most images of the CTCUG dataset, the background is in focus and the foreground is blurry. More detailed descriptions about the CTCUG dataset can be found in [27]. Thus, in this paper, we adopt a single supervision method after our comprehensive consideration.

## 5 Conclusion

In this paper, we propose a DBD method based on transformer encoder and edge guidance, which aims to detect low-contrast regions and distinguish the boundaries of in-focus objects. First, We utilize a transformer encoder to capture the global context information, whic- h helps to detect low-contrast regions and suppress the background clutter. We also deploy CNNs to model local detailed features to perform better in localization. Therefore, a hybrid CNN-Transformer architecture is adopted from a top-bottom manner as our backbone. Second, we develop an EENet to obtain local edge information of in-focus objects. Additionally, EGFNet can effectively combine local edge information with global semantic features to produce the fused features with accurate boundaries. Finally, as a decoder, HFANet can further hierarchically decode and refine the feature maps with clear edges. Experimental results demonstrate that our model outperforms state-of-the-art methods on three datasets without any pre-processing or post-processing.

Our method could occasionally fail for detecting the object boundaries of large low contrast regions. As shown in Fig. 12, this is due to the fact that there is no strong edge within such regions for extraction and fusion. For future works, we will use a transformer encoder to solve the detection of boundary information.

**Funding** 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).

**Availability of data and material** The datasets used or analysed during the current study are available from the corresponding author on reasonable request.

**Code Availability** The code of our model can be obtained from the corresponding author on reasonable request, and the code will be released at https://github.com/zzjssr/TransDBD.

## Declarations

**Conflict of Interests** We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

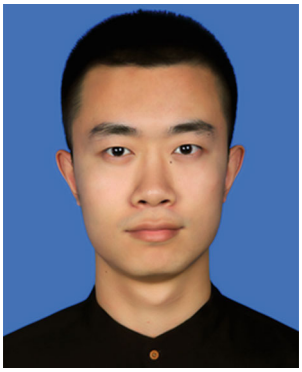
## References

1. Tang C, Hou C, Song Z (2013) Defocus map estimation from a single image via spectrum contrast[J]. Opt Lett 38(10):1706–1708. https://doi.org/10.1364/OL.38.001706

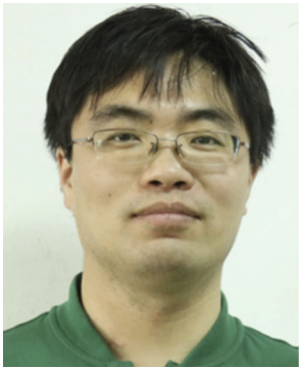
2. Xia C, Gao X, Li KC et al (2020) Salient object detection based on distribution-edge guidance and iterative Bayesian optimization. Appl Intell 50:2977–2990. https://doi.org/10.1007/s10489-020-01691-7
3. Levin A, Rav-Acha A, Lischinski D (2008) Spectral matting[J]. IEEE Trans Pattern Anal Machine Intell 30(10):1699–1712. https://doi.org/10.1109/TPAMI.2008.168
4. Zhang X, Wang R, Jiang X et al (2016) Spatially variant defocus blur map estimation and deblurring from a single image[J]. J Visual Commun Image Represent 35(Feb):257–264. https://doi.org/10.1016/j.jvcir.2016.01.002
5. Zhu X, Cohen S, Schiller S et al (2013) Estimating spatially varying defocus blur from a single image[J]. IEEE Trans Image Process 22(12):4879–4891. https://doi.org/10.1109/TIP.2013.2279316
6. Vu CT, Phan TD, Chandler DM (2012) $S_3$: A spectral and spatial measure of local perceived sharpness in natural images. In: IEEE Transactions on Image Processing, pp 934–945. https://doi.org/10.1109/TIP.2011.2169974
7. Zhang Y, Hirakawa K (2013) Blur processing using double discrete wavelet transform. In: IEEE Conference on computer vision and pattern recognition, pp 1091–1098. https://doi.org/10.1109/CVPR.2013.145
8. Shi J, Xu L, Jia J (2014) Discriminative blur detection features. In: IEEE conference on computer vision and pattern recognition, pp 2965–2972. https://doi.org/10.1109/CVPR.2014.379
9. 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 Processing Letters, 1652–1656. https://doi.org/10.1109/LSP.2016.2611608
10. Park J, Tai Y, Cho D et al (2017) A unified approach of multi-scale deep and hand-crafted features for defocus estimation. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2760–2769. https://doi.org/10.1109/CVPR.2017.295
11. Zhuo S, Sim T (2011) Defocus map estimation from a single image[J]. Pattern Recogn 44(9):1852–1858. https://doi.org/10.1016/j.patcog.2011.03.009
12. Zhao J, Feng H, Xu Z et al (2013) Automatic blur region segmentation approach using image matting[J]. SIViP 7(6):1173–1181. https://doi.org/10.1007/s11760-012-0381-6
13. Su B, Lu S, Tan Ch L (2011) Blurred image region detection and classification. In: ACM International Conference on Multimedia, pp 1397–1400
14. Saad E, Hirakawa K (2016) Defocus blur-invariant scale-space feature extractions[J]. IEEE Trans Image Process 25(7):3141–3156. https://doi.org/10.1109/TIP.2016.2555702
15. Pang Y, Zhu H, Li X et al (2017) Classifying discriminative features for blur detection[J]. IEEE Trans Cybern 46(10):2220–2227. https://doi.org/10.1109/TCYB.2015.2472478
16. Liu R, Li Z, Jia J (2008) Image partial blur detection and classification. In: IEEE conference on computer vision and pattern recognition, pp 1–8. https://doi.org/10.1109/CVPR.2008.4587465
17. 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
18. 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, 3142–3155. https://doi.org/10.1109/TIP.2017.2662206
19. Dong C, Loy CC, He K et al (2016) Image super-resolution using deep convolutional networks[J]. IEEE Trans Pattern Anal Mach Intell 38(2):295–307. https://doi.org/10.1109/TPAMI.2015.2439281
20. Li P, Wang D, Wang L et al (2018) Deep visual tracking: review and experimental comparison. Pattern Recogn, 323–338
21. Wei Y, Wei X, Min L et al (2016) HCP: A flexible cnn framework for multi-label image classification[J]. IEEE Trans Softw Eng 38(9):1901–1907. https://doi.org/10.1109/TPAMI.2015.2491929
22. Park J, Tai YW, Cho D et al (2017) A unified approach of multi-scale deep and hand-crafted features for defocus estimation[J]. IEEE Computer Society, 2760–2769
23. Zhao W, Zhao F, Wang D et al (2018) Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network. In: IEEE conference on computer vision and pattern recognition, pp 3080–3088. https://doi.org/10.1109/CVPR.2018.00325
24. Zhao W, Zheng B, Lin Q et al (2019) Enhancing diversity of defocus blur detectors via cross-ensemble network. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 8897–8905. https://doi.org/10.1109/CVPR.2019.00911
25. Ma K, Fu H, Liu T et al (2016) Deep blur mapping: exploiting high-level semantics by deep neural networks[J]. IEEE Trans Image Process 5155-5166:27. https://doi.org/10.1109/TIP.2018.2847421
26. Lee J, Lee S, Cho S et al (2019) Deep defocus map estimation using domain adaptation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 12214–12222. https://doi.org/10.1109/CVPR.2019.01250
27. 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 Machine Intell PP(99):1–1. https://doi.org/10.1109/TPAMI.2020.3014629
28. Tang C, Zhu X, Liu X et al (2019) DefusionNET: defocus blur detection via recurrently fusing and refining multi-scale deep features. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 2695–2704. https://doi.org/10.1109/CVPR.2019.00281
29. Tang C, Liu X, An S et al (2021) BR$^2$Net: defocus blur detection via a bidirectional channel attention residual refining network. IEEE Transactions on Multimedia, 624–635. https://doi.org/10.1109/TMM.2020.2985541
30. Tang C, Liu X, Zhu X et al (2020) R$^2$MRF: defocus blur detection via recurrently refining multi-scale residual features[J]. Proc AAAI Conf Artif Intell 34(7):12063–12070. https://doi.org/10.1609/aaai.v34i07.6884
31. Li J, Fan D, Yang L et al (2021) Layer-output guided complementary attention learning for image defocus blur detection. IEEE Trans Image Process, 3748–3763. https://doi.org/10.1109/TIP.2021.3065171
32. Hu J, Shen L, Albanie S et al (2017) Squeeze-and-excitation networks. In: IEEE transactions on pattern analysis and machine intelligence, pp 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372
33. Peng C, Zhang X, Yu G et al (2017) Large kernel matters — improve semantic segmentation by global convolutional network. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 1743–1751. https://doi.org/10.1109/CVPR.2017.189
34. Zhao J, Liu J, Fan D et al (2020) EGNet: edge guidance network for salient object detection. In: IEEE/CVF international conference on computer vision (ICCV), pp 8778–8787. https://doi.org/10.1109/ICCV.2019.00887
35. Chen Z, Xu Q, Cong R et al (2020) Global context-aware progressive aggregation network for salient object detection[J]. Proc AAAI Conf Artif Intell 34(7):10599–10606. https://doi.org/10.1609/aaai.v34i07.6633
36. Wei J, Wang S, Huang Q (2019) F3Net: fusion, feedback and focus for salient object detection[J]. arXiv:1911.11445
37. Fan D, Cheng M, Liu Y et al (2017) Structure-measure: a new way to evaluate foreground maps. In: IEEE international conference on computer vision (ICCV), pp 4558–4567. https://doi.org/10.1109/ICCV.2017.487

38. 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
39. Xin Y, Eramian M (2016) LBP-Based segmentation of defocus blur[J]. IEEE Trans Image Process 25(4):1–1. https://doi.org/10.1109/TIP.2016.2528042
40. Tang C, Wu J, Hou Y et al (2016) A spectral and spatial approach of coarse-to-fine blurred image region detection[J]. IEEE Signal Process Lett 23(11):1652–1656. https://doi.org/10.1109/LSP.2016.2611608
41. Golestaneh SA, Karam LJ (2017) Spatially-varying blur detection based on multiscale fused and sorted transform coefficients of gradient magnitudes. In: IEEE conference on computer vision and pattern recognition, pp 5800–5809. https://doi.org/10.1109/CVPR.2017.71
42. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778. https://doi.org/10.1109/CVPR.2016.90
43. Dosovitskiy A, Beyer L, Kolesnikov A et al (2021) An image is worth 16×16 words: Transformers for image recognition at scale. In: Proceedings of the the international conference on learning representations (ICLR)
44. Zheng S, Lu J, Zhao H et al (2021) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers
45. Wang W, Xie E, Li X et al (2021) Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. arXiv:2102.12122
46. Ranftl R, Bochkovskiy A, Koltun V (2021) Vision transformers for dense prediction. arXiv:2103.13413
47. Chen J, Lu Y, Yu Q et al (2021) TransUNet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306
48. Mao Y, Zhang J, Wan Z et al (2021) Transformer transforms salient object detection and camouflaged object detection. arXiv:2104.10127
49. Sun P, Jiang Y, Zhang R et al (2020) TransTrack: multiple-object tracking with transformer. arXiv:2012.15460





**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.