Self-generated Defocus Blur Detection via Dual Adversarial Discriminators

Wenda Zhao[1]*, Cai Shang[1], Huchuan Lu[1]
[1] School of Information and Communication Engineering, Dalian University of Technology, China
zhaowenda@dlut.edu.cn; 1147626517@dlut.mail.edu.cn; lhchuan@dlut.edu.cn

Abstract

Although existing fully-supervised defocus blur detection (DBD) models significantly improve performance, training such deep models requires abundant pixel-level manual annotation, which is highly time-consuming and error-prone. Addressing this issue, this paper makes an effort to train a deep DBD model without using any pixel-level annotation. The core insight is that a defocus blur region/focused clear area can be arbitrarily pasted to a given realistic full blurred image/full clear image without affecting the judgment of the full blurred image/full clear image. Specifically, we train a generator G in an adversarial manner against dual discriminators D_c and D_b. G learns to produce a DBD mask that generates a composite clear image and a composite blurred image through copying the focused area and unfocused region from corresponding source image to another full clear image and full blurred image. Then, D_c and D_b can not distinguish them from realistic full clear image and full blurred image simultaneously, achieving a self-generated DBD by an implicit manner to define what a defocus blur area is. Besides, we propose a bilateral triplet-excavating constraint to avoid the degenerate problem caused by the case one discriminator defeats the other one. Comprehensive experiments on two widely-used DBD datasets demonstrate the superiority of the proposed approach. Source codes are available at: ***https://github.com/shangcai1/SG****.*

1. Introduction

Defocus blur will emerge when the scene is out of the camera's focus distance, which is a common phenomenon in an image. Defocus blur detection (DBD) can be potentially used to many vision tasks (*e.g.*, salient region detection [9], autofocus [39], depth estimation [19, 5]). Thus, DBD has been gaining more and more research interest. Recently, deep convolutional neural networks (CNNs)-based DBD methods [25, 24, 43, 27, 44, 25, 11, 38, 45] achieve a high performance through abundant labeled data (*e.g.*, pixel-level annotation [20, 43]). However, manual annotation is time-consuming and error-prone. In order to bridge this gap, we make an effort to obtain DBD directly from real images without using any pixel-level annotation.

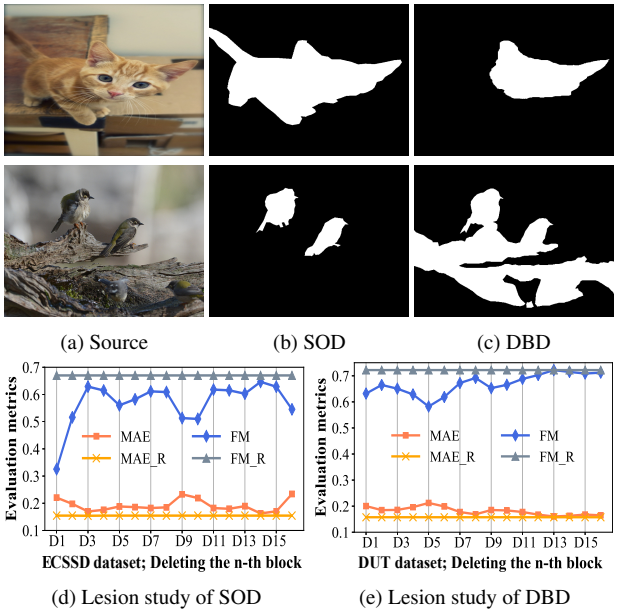


Figure 1. Semantic correlation comparison of saliency object detection (SOD) and defocus blur detection (DBD). (a)-(c): source image, SOD result, and DBD result. (d) and (e) are lesion studies of SOD and DBD on ECSSD dataset [33] and DUT dataset [44], respectively. ResNet50 [7] is fully trained, and then we measure the importance of each convolution block by removing it, leaving the skip connection unchanged.

Intuitively, some unsupervised segmentation tasks, such as saliency object detection [35, 36] and semantic segmentation [31, 14, 8, 13, 30, 17], can be used to relieve our problem. For example, Zhang *et al.* [36] propose a "supervision by fusion" strategy through generating reliable supervisory signals in the process of weak saliency model fusion. Bielski *et al.* [2] adopt the idea, where objects can be moved locally independently of a given background, to design perturbed generative models for unsupervised object segmentation. In brief, these methods always employ ob-

*Corresponding author: Wenda Zhao (zhaowenda@dlut.edu.cn).

ject semantic information to achieve unsupervised segmentation. However, the object semantic information is weakly related to DBD [37]. Figures 1 (a)-(c) show the visual comparison. In contrast to saliency object detection that segments semantic objects (*e.g.*, cat and bird), DBD detects focused clear areas that ignores semantic integrity. For example, the clear head of a cat and partial weak semantic stumps are detected. Moreover, inspired by the validation in [28] that residual networks can be seen as a collection of many paths where they do not strongly depend on each other, we delete individual convolution block path from ResNet50 [7] after it has been fully trained to analysis the semantic correlation. As shown in Figure 1 (e), deleting any high-level semantic blocks (D12-D15) has no a noticeable performance change. However, deleting some low-level convolution blocks (D1-D11) will reduce the performance. This is contrary to saliency object detection that strongly relies on high-level semantic features (see Figure 1 (d)). Therefore, weak semantic correlation brings a larger challenge for unsupervised DBD.

Based on the attribute of weak semantic correlation in DBD, a principle can be acquired: A defocus blur region can be arbitrarily moved relative to a given realistic full blurred image without affecting the judgment of the full blurred image; Similarly, a focused clear region can be randomly pasted to a given realistic full clear image without affecting the criterion of the full clear image. Sequentially, we propose a unsupervised learning framework [1] through this principle.

The core idea is that we firstly build a generative network to output a DBD mask without ground truth as supervision. Then, the focused and unfocused regions are cut out through the predicted mask from corresponding source image, respectively. Afterwards, two composite images of C_c and C_b, obtained by pasting the focused region and unfocused region to another full clear image and full blurred image, are fed into two discriminative networks, respectively. The discriminators aim to distinguish whether C_c is a realistic full clear image and C_b is a realistic full blurred image. In order to fool the discriminators to believe that, the generator must output a DBD mask that accurately cuts out the focused region and unfocused area from corresponding source image. Therefore, we achieve an implicit manner to define what a defocus blur area is, that avoids manual labelling.

Especially, the motivation of implementing dual adversarial discriminative networks is to avoid a degenerate solution that generates a partial or an excessive DBD mask to fool the discriminator successfully. Specifically, if a single discriminative network is adopted, a partial DBD mask can generate a full clear image to fool the discriminator (see the second row of Figure 2). On the other hand, an excessive DBD mask can generate a full blurred image to fool another discriminator (see the third row of Figure 2). Our proposed dual adversarial discriminative networks can alleviate this problem. Since a partial DBD mask can generate a full clear image to fool one discriminator, but can not simultaneously generate a full blurred image to fool another discriminator. Similarly, the principle is used to avoid the failure of generating an excessive DBD mask. Therefore, only if the produced DBD mask is accurate can the composite blurred image and composite clear image fool dual discriminators successfully (see the fourth row of Figure 2).

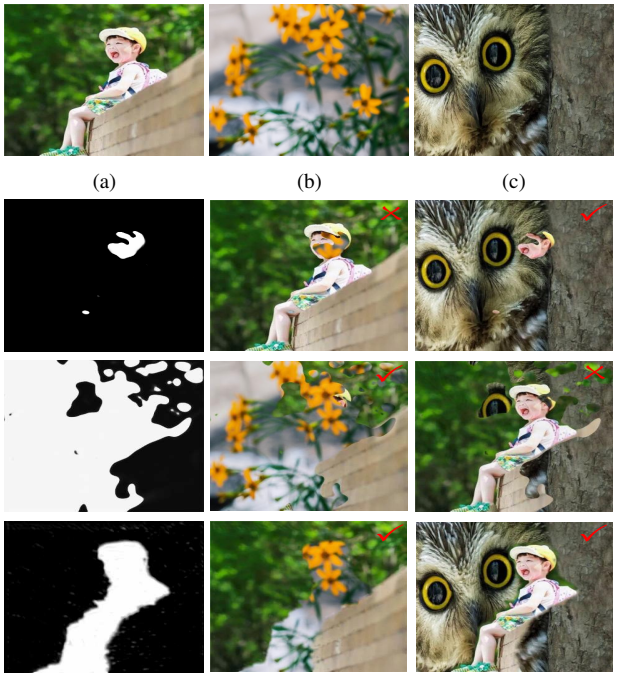


Figure 2. Motivation of implementing dual adversarial discriminative networks. (a)-(c) are source image, full clear image, and full blurred image, respectively. (d)-(f) are DBD mask, composite blurred image, and composite clear image, respectively. In the second row, a partial DBD mask can compose a full clear image to fool one discriminator (expressed as ✓), but can not generate a full blurred image to fool another discriminator (marked as ×). An excessive DBD mask has the opposite effect (see the third row). In contrast, the DBD mask generated with the help of dual adversarial discriminative networks is accurate, such that the composite blurred image and composite clear image can fool dual discriminators successfully, as shown in the fourth row.

Additionally, in the adversarial training process for our unsupervised framework, one discriminator easily defeats the other one, which will force the generator to generate a full DBD mask or an empty DBD mask. To avoid this failure, we propose a bilateral triplet-excavating constraint to effectively balance these two discriminators. Specifically, we first implement a classification network to excavate the feature relationship of triplet images among the realistic

[1] Here, "unsupervised" means that our method achieves DBD without using any pixel-level manual annotation.

full clear image, full blurred image and mixed image which includes both clear region and blurred area. Afterwards, we encourage the feature-space distance between the composite clear image C_c and another realistic full clear image to get closer, and simultaneously inspire the distance of the composite blurred image C_b and another realistic full blurred image to be smaller. With this constraint, the two discriminators can easily achieve a balance, which thereby makes the mechanism of encouraging the generator to produce an accurate DBD mask to simultaneously fool the two discriminators come into force.

Main contributions in this paper are summarized as follows.

- We make an effort to train an effective deep defocus blur detector without using any pixel-level manual annotation.
- We build dual adversarial discriminative networks to force the generator to produce an accurate DBD mask.
- We propose a bilateral triplet-excavating constraint to avoid the degenerate problem, where one discriminator defeats the other one, easily making the generator produce a full or an empty DBD mask.

We validate the effectiveness of the proposed unsupervised module on two widely-used benchmark datasets.

2. Related Work

2.1. Fully Supervised DBD Methods

Recently, CNN-based methods [43, 25, 44, 11, 27, 38, 12, 45, 25] have substantially improved the performance of DBD. Among these methods, multi-level feature integration strategy is mainly studied. For instance, Kim *et al.* [11] implement a encoder-decoder network with long residual skip-connections to combine low-level structural features and high-level contextual features. Tang *et al.* [27] fuse and refine the multi-level features by a cross-layer manner, where low-level features are propagated to top layers for refining the details and high-level features are propagated to bottom layers to help locate the defocus regions. Besides, some other strategies are successfully implemented for DBD. Zhao *et al.* [43] propose a multi-stream bottom-top-bottom network to fuse multi-scale image information. Zhang *et al.* [38] use a dilated convolution with pyramid pooling to preserve details. Cross-ensemble network [45], which enhances diversity of multiple DBD learners, is designed to improve accuracy and speed. In particular, Zhang *et al.* [37] adopt a cut-and-paste scheme for data augmentation and combine pixel-level fully supervised learning for DBD.

However, training these models requires abundant expensive pixel-level labels. Addressing this problem, we propose a unsupervised learning framework to obtain DBD mask directly from the collection of real images without any pixel-level annotation. Our method is the first attempt to define what a defocus blur region is automatically, through identifying a powerful principle based on the characteristic of the weak semantic correlation of DBD.

2.2. Unsupervised DBD Methods

Unsupervised methods for DBD mostly utilize hand-crafted features in transform domain [40, 21] and spatial domain [32, 10, 15] to measure defocus blur. For example, Shi *et al.* [21] use sparse representation and image decomposition to extract blur features. Golestaneh *et al.* [1] propose a sort transform of gradient magnitude and a high frequency multiscale fusion to estimate defocus blur. Yi *et al.* [34] adopt the local binary patterns to design a sharpness metric. Besides, the feature integration of the spatial domain and transform domain is investigated [19, 26]. For instance, Shi *et al.* [20] combine gradients, Fourier features and local filter features to detect blur. Besides, Zhao *et al.* [42] design a weakly-supervised recurrent constraint network for DBD using bounding box annotation.

Hand-crafted-based unsupervised methods and weakly-supervised method have been demonstrated to be effective in some specific cases. However, they are not robust to distinguish blur for complex scenes.

2.3. Unsupervised Object Segmentation Methods

Unsupervised methods have been explored in many visual tasks, *e.g.*, person re-identification [29], image restoration [3], and image fusion [41]. The work that closely relates to ours is unsupervised object segmentation. On one hand, unsupervised domain adaptation, where the models trained from synthetic data transfer to unlabeled source images, is explored to mainly address domain shift problem for object segmentation. Wang *et al.* [31] ease the domain shift between the synthetic data and the real data to improve semantic-level alignment. Pan *et al.* [14] propose a two-step self-supervised domain adaptation approach to minimize the inter-domain and intra-domain gap. On the other hand, semantic correlation within a object is utilized to build unsupervised object segmentation framework. [13, 30] excavate the inherent correlation among video frames and design attention mechanisms to obtain object segmentation without using labels as supervision. Prior knowledge of the object (*e.g.*, shape and contrast) is adopted to structure an unsupervised method in [6]. In addition, [17, 2] build on the idea that objects can move independently of their background to achieve the unsupervised object segmentation.

Existing CNN-based unsupervised object segmentation methods mainly take advantage of the object semantic correlation. However, the semantic information is weakly re-

lated to DBD. Thus, the ideas of unsupervised object segmentation methods can hardly be used to our task. In this paper, we capture a principle that a defocus blur region or focused clear area can move without affecting the judgment of its blur or clear region, and propose dual adversarial discriminative networks to force a generator to produce a accurate DBD mask without using any pixel-level manual annotation. Besides, we explore the degenerate problem, where one discriminator defeats the other one to make the generator generate a full or empty mask, and further propose a bilateral triplet-excavating constraint to avoid it.

3. Learning to Detect Defocus Blur without Pixel-level Annotation

Our framework builds on the successful approach of generative adversarial network (GAN) [4], which includes two main building blocks: A generator G and two discriminators D_c and D_b, as shown in Figure 3. G is trained against D_c and D_b in an adversarial manner, *i.e.*, G learns to produce a DBD mask that generates two composite images of C_c and C_b through copying the focused area and unfocused region from corresponding source image to another full clear image and full blurred image, and then D_c and D_b can not distinguish them from realistic full clear image and full blurred image simultaneously.

One aspect that we would like to highlight: We introduce random realistic full clear image and full blurred image unknown to the generator G in the training process. Therefore, the generator G must output a DBD mask that can combine corresponding focused and unfocused regions with arbitrary full clear image and full blurred image to fool two discriminators D_c and D_b simultaneously into believing that the composite images are full clear and full blurred. This is an implicit manner to define what a defocus blur area is, without using any manual annotation.

3.1. A Generator for DBD Mask

Consider N training samples $\mathbf{T} = \{T_1, T_2, ..., T_N\}$, and T_n has $M = H \times W \times K$ dimensional, where H, W and K stand for the height, width and number of channels of the n-th sample, respectively. Note that the training samples have no corresponding manual pixel-level annotation. We aim to train a generative model taking the form: $M_n = G(T_n; W_G)$ that predicts a DBD mask $M_n \in [0, 1]$ given an image T_n, where W_G is weight parameters of the generative model.

We would like to generate two composite images of C_c and C_b by utilizing the mask M_n to move the focused and unfocused regions from corresponding source image to another full clear image I_c and full blurred image I_b, respectively. Formally, this can be written as

$$C_c(M_n) = M_n \otimes T_n \oplus (I - M_n) \otimes I_c, \tag{1}$$

$$C_b(M_n) = (I - M_n) \otimes T_n \oplus M_n \otimes I_b, \tag{2}$$

where $\otimes$ stands for a pixel-wise multiplication operation, $\oplus$ expresses pixel-wise addition operation, and I indicates a matrix in which all elements are 1. Visual description is shown in the grey path of Figure 3.

Then, the generator G is trained in an adversarial manner against two discriminators D_c and D_b (see the orange path in Figure 3), where we minimize the following loss

$$\begin{aligned}\mathcal{L}_G(W_G) = E_{\mathbf{T}\sim P_t}[log(1 - D_c(C_c(G(\mathbf{T}; W_G))))] \\ + E_{\mathbf{T}\sim P_t}[log(1 - D_b(C_b(G(\mathbf{T}; W_G))))],\end{aligned} \tag{3}$$

where W_G expresses weight parameters of the generator G, and P_t illustrates the probability density distribution of training images.

3.2. Dual Adversarial Discriminators

If a single discriminator D_c or D_b is implemented to distinguish whether C_c or C_b is a full clear image or a full blurred image, a trivial solution will be created that the generator produces a partial or an excessive DBD mask to fool the discriminator successfully (see Figure 2). Here, dual adversarial discriminators are designed to alleviate this problem. A partial DBD mask can generate a full clear image to fool the discriminator D_c, but can not simultaneously generate a full blurred image (*i.e.*, the generated image contains partial clear regions) to fool the discriminator D_b. Vice versa, an excessive DBD mask can not generate a full clear image to fool the discriminator D_c. To train the dual discriminators, we maximize the following loss

$$\begin{aligned}\mathcal{L}_D(W'_D; W''_D) = E_{\mathbf{I}_c\sim P_c}[log(D_c(\mathbf{I}_c; W'_D))] \\ + E_{\mathbf{T}\sim P_t}[log(1 - D_c(C_c(G(\mathbf{T}; W_G))))] \\ + E_{\mathbf{I}_b\sim P_b}[log(D_b(\mathbf{I}_b; W''_D))] \\ + E_{\mathbf{T}\sim P_t}[log(1 - D_b(C_b(G(\mathbf{T}; W_G))))],\end{aligned} \tag{4}$$

where W'_D and W''_D are weight parameters of the discriminators D_c and D_b, respectively. P_c and P_b stand for the probability density distributions of realistic full clear images and full blurred images, respectively. $\mathbf{I}_c$ and $\mathbf{I}_b$ are full clear training sample dataset and full blurred training sample dataset, respectively. Illustration to this process is shown in the orange path of Figure 3.

Now, the generator has an incentive to produce a accurate DBD mask, since a partial or an excessive DBD mask will synthesize a mixed image, which includes both clear region and blurred area. And the discriminators can successfully distinguish it.

3.3. Averting A Degenerate Solution

Our unsupervised framework aims to implement dual adversarial discriminators, forcing the generator to generate a accurate DBD mask, where a degenerate case easily

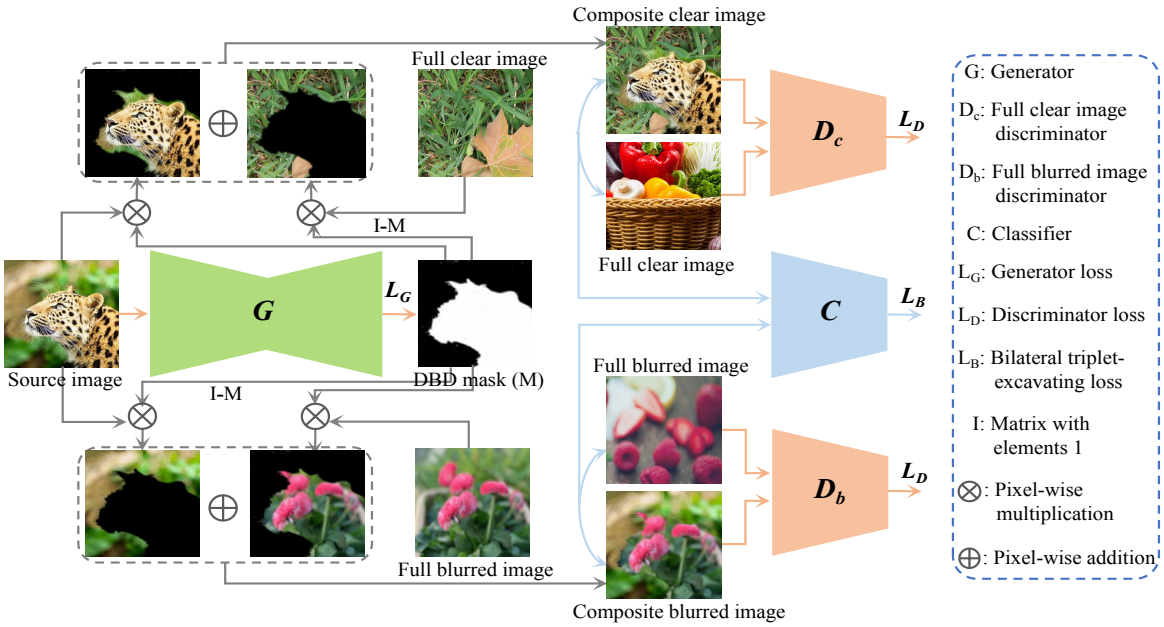


Figure 3. Architecture illustration of the proposed framework that generates DBD without using pixel-level annotation. The framework consists of three components: (1) Dual adversarial discriminative models are designed that force the generator G to produce DBD mask M accurately (**orange path**); (2) Composite clear image and composite blurred image are generated through utilizing M to paste the clear area and blurred region of source image to a full clear image and full blurred image respectively (**grey path**); (3) A bilateral triplet-excavating constraint model is built by a classifier to balance the dual discriminators (**blue path**).

emerges in the adversarial training process, *e.g.*, one discriminator wins the other one that makes the the generator produce a full DBD mask or an empty DBD mask. Specifically, if the discriminator D_c defeats the discriminator D_b, the generator G will produce an empty DBD mask to select a realistic full clear image as the synthesized image, and vice versa.

This case can be automatically relieved in our framework, since the generator is encouraged to fool the two discriminators simultaneously. Interestingly, we find that adding a bilateral triplet-excavating constraint loss significantly assists a balance of the two discriminators and improves DBD performance. Specifically, we encourage the feature-space distance between the composite clear image C_c and another realistic full clear image to get closer, and simultaneously inspire the feature-space distance between the composite blurred image C_b and another realistic full blurred image to be smaller. This gives a bilateral triplet-excavating loss to the generator

$$\mathcal{L}_B(W_G) = 2 - \{cos[\frac{|V(C_c(M_n))V(I_c(M_n))|}{|V(C_c(M_n))||V(I_c(M_n))|}] + cos[\frac{|V(C_b(M_n))V(I_b(M_n))|}{|V(C_b(M_n))||V(I_b(M_n))|}]\}. \quad (5)$$

Here, we implement a classification network to excavate the triplet relationship among the realistic full clear image, full blurred image and mixed image. This can ensure that the feature $V(\cdot)$ is related to defocus blur.

3.4. Joint Training

Summarizing the above descriptions, our overall loss function is

$$\mathcal{L}_O(W_G; W'_D; W''_D) = \eta\mathcal{L}_G(W_G) + \mathcal{L}_D(W'_D; W''_D) + \mathcal{L}_B(W_G), \quad (6)$$

where η is a hyperparameter. Therefore, together with the bilateral triplet-excavating loss, the generator is trained in an adversarial manner against two discriminators, achieving a unsupervised framework for DBD without using any manual pixel-level annotation successfully.

3.5. Architecture

As shown in Figure 3, our unsupervised framework consists of three modules: (1) A generator G, which is used to generate DBD mask; (2) Two discriminators D_c and D_b, which are implemented to distinguish whether the composite images C_c and C_b are a full clear image and a full blurred image, respectively; (3) A classification network, which is employed to excavate the triplet relationship among the realistic full clear image, full blurred image and mixed image. We briefly describe their architectures as follows.

Generator. The structure of our generator is similar to U-Net [18]. Specifically, the first five convolution blocks (denoted as CB_1, CB_2, CB_3, CB_4, CB_5) of VGG16 [22] is utilized as the encoder to extract high-level features. Correspondingly, a decoder consisting of 4 convolution blocks (named as DCB_1, DCB_2, DCB_3, DCB_4)

Table 1. Importance study of dual adversarial discriminators using F-measure and MAE values on both DUT and CUHK datasets. D_c and D_b are the discriminators that distinguish whether the composite image is full clear and full blurred, respectively.

Scheme	CUHK		DUT	
	F-measure	MAE	F-measure	MAE
Single D_c	0.360	0.264	0.372	0.282
Single D_b	0.353	0.699	0.367	0.686
D_c and D_b	**0.719**	**0.148**	**0.683**	**0.190**

generates the DBD mask, which has the same resolution with the source image. In front of DCB_x, a bilinear interpolation is used to upsample the features, and a skip connection from CB_x ($x = 1, 2, 3, 4$) is added after $DCB\_(5-x)$ to integrate low-level detailed features.

Discriminators. Discriminators D_c and D_b have the same architecture, where 9 convolutions with kernel size of 3×3 are stacked to extract high-level features. Then, a global average pooling operation is adopted to reduce dimension as a vector of 512×1, and followed by two fully connected layers FC_1024 and FC_1 to generate a one-dimensional vector, which judges whether the composite image C_c is full clear or C_b is full blurred.

Classifier. The classifier uses the first five convolution blocks of VGG16 to extract high-level features. After that, a global average pooling operation is implemented to reduce dimension as a vector of 512×1, and then two fully connected layers FC_128 and FC_3 to generate a three-dimensional vector, which discriminates whether the input image is a full clear image, full blurred image or mixed image. Here, the vector of 512×1 is adopted to align the features of the realistic full-clear image I_c with the composite full-clear image C_c, and the realistic full-blur image I_b with the composite full-blur image C_b through Eq. (5).

4. Experiments

4.1. Experimental Setup

Datasets. Two common datasets CUHK [20] and DUT [43, 44] with pixel-level annotation are adopted in this work. We implement the same strategy with [44], where the number of training images & testing images are 604 & 100 in CUHK dataset and 600 & 500 in DUT dataset respectively, to evaluate our method. Notice that, pixel-level annotation is not used in the process of training our model.

In addition, we construct a new dataset (FCFB) including 500 natural full clear images and 500 natural full blurred images to facilitate the training of our model. FCFB contains a variety of unfocused blurred and focused clear scenes, making our model be successfully implemented.

Implementation. We implement the proposed model using Pytorch on a RTX 2080Ti GPU with batch size 4. Adam with learning rate 0.0002 and momentum 0.9 is utilized as the optimizer. Classifier is pretrained to obtain the capability of excavating the triplet relationship among the full clear image, full blurred image and mixed image. Then, we train the generator and dual adversarial discriminators to produce a DBD mask without using manual pixel-level annotation.

Table 2. Influence study of bilateral triplet-excavating constraint through adjusting η.

Weight	CUHK		DUT	
	F-measure	MAE	F-measure	MAE
η=0.000	0.719	0.148	0.683	0.190
η=0.005	**0.778**	**0.118**	0.686	0.183
η=0.010	0.769	0.119	**0.701**	**0.172**
η=0.050	0.738	0.133	0.673	0.182
η=0.500	0.734	0.148	0.670	0.197

Table 4. Effect study of dataset FCFB using F-measure and MAE values on both DUT and CUHK datasets. FCSB(nD) stands for simulated dataset containing n blur-degree images and natural full clear images.

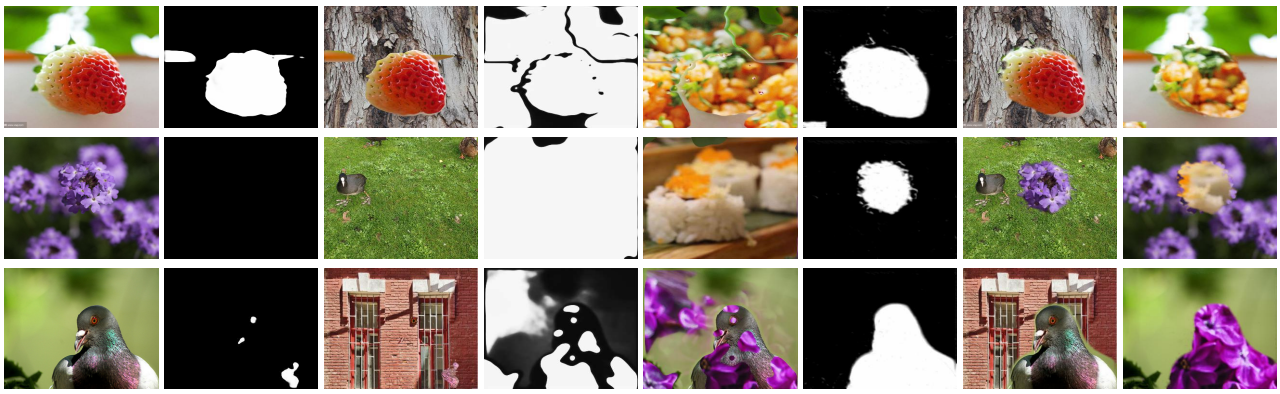
Training dataset	CUHK		DUT	
	F-measure	MAE	F-measure	MAE
FCSB(1D)	0.715	0.150	0.610	0.215
FCSB(2D)	0.746	0.141	0.621	0.213
FCSB(3D)	0.760	0.132	0.639	0.208
FCSB(4D)	0.757	0.130	0.624	0.208
FCSB(5D)	0.768	0.123	0.655	0.192
FCFB	**0.769**	**0.119**	**0.701**	**0.172**

Evaluation. We adopt four evaluation methods including F-measure value [43, 44], mean absolute error (MAE), F-measure curve and Precision-Recall (PR) curve to evaluate the performance of the proposed model. The F-measure is an overall performance measurement, which is calculated as follow: $F_\beta = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}$, where β^2 is 0.3. A larger F_β indicates a better performance. MAE is used to measure the pixel-wise dissimilarity between the ground truth G_t and DBD mask M, which is defined as: $MAE = \frac{1}{P} \sum_{p=1}^{P} |G_t(p) - M(p)|$, where p stands for the pixel position, and P is the total number of pixels. A smaller MAE demonstrates a more accurate result.

4.2. Ablation Study

Importance of Dual Adversarial Discriminators. In Sec. 3.2, we introduce dual adversarial discriminators D_c and D_b that force the generator G to generate an accurate DBD mask. Here, we compare two schemes. One is using a single discriminator D_c to make G produce a DBD mask, such that generates a composite clear image to fool D_c. The other one is utilizing a single discriminator D_b to force G to generate a DBD mask, such that obtains a composite blurred image to fool D_b. The results are shown in Table 1. Dual adversarial discriminators can achieve significant performance improvement.

Visual comparison is shown in Figure 4. When a sin-



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

Figure 4. Visual comparison of DBD masks generated by different schemes. (a) source; (b)-(c) DBD mask with single discriminator D_c and corresponding composite clear image; (d)-(e) DBD mask with single discriminator D_b and corresponding composite blurred image; (f)-(h) DBD mask with dual discriminators D_c and D_b, and corresponding composite clear image and blurred image. A single discriminator D_c or D_b can be successfully cheated by a degenerate solution. However, only the DBD mask is accurate that can cheat dual discriminators D_c and D_b simultaneously.

Table 3. Quantitative comparison of the state-of-the-art approaches using F-measure and MAE scores. "(U)" and "(S)" denote unsupervised and supervised with pixel-level manual annotation, respectively. Average time is calculated on a workstation with a RTX 2080Ti 11G GPU for an image size 320×320.

	Metric	Unsupervised methods								Supervised methods	
		SVD [23]	DBDF [20]	SRID [21]	KSFV [15]	SS [26]	DHCF [16]	HiFST [1]	Ours(U)	Ours(S)	BTBNet [43]
DUT	F-measure	0.664	0.503	0.493	0.562	0.669	0.471	0.686	**0.701**	0.794	0.767
	MAE	0.282	0.376	0.516	0.271	0.293	0.412	0.251	**0.172**	0.153	0.197
CUHK	F-measure	0.750	0.579	0.446	0.521	0.701	0.477	0.701	**0.769**	0.884	0.861
	MAE	0.242	0.311	0.573	0.300	0.270	0.374	0.233	**0.119**	0.079	0.113
Average time	Seconds	7.558	45.45	3.758	19.13	0.395	11.76	47.62	**0.005**	0.005	25.00

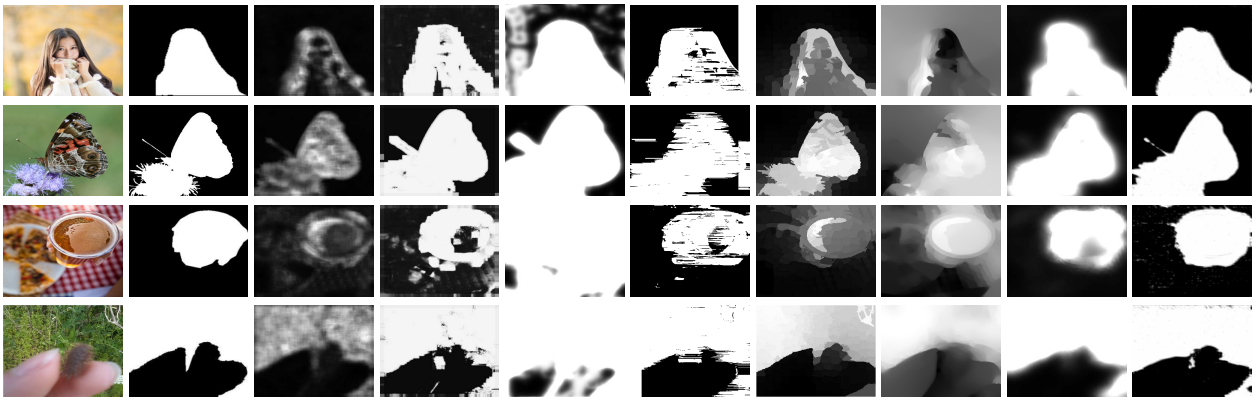
gle discriminator D_c or D_b is used, a degenerate solution can generate a composite clear image or blurred image that successfully cheats the discriminator, as shown in Figures 4 (b)-(e). In contrast, only dual discriminators D_c and D_b are adopted to produce an accurate mask, which can make the composite clear image and blurred image cheat D_c and D_b simultaneously, as shown in Figures 4 (f)-(h).

Influence of Bilateral Triplet-excavating Constraint. In Sec. 3.3, we propose a bilateral triplet-excavating constraint to assist a balance of the dual discriminators in the process of adversarial training. We study the influence of the proposed bilateral triplet-excavating constraint by relatively adjusting the parameter η in Eq. (6). Table 2 shows the results. With the increase of η, the performance first becomes better and then decreases. The reason is that a small η makes bilateral triplet-excavating constraint dominant, and a large η will relatively restrain the effect of dual adversarial discriminators. Here, we take $\eta = 0.01$ for experiments.

Effect of Dataset FCFB. We collect a new dataset FCFB consisting of 500 natural full clear images and 500 natural full blurred images to help facilitate the training of our model. We study effect of FCFB by comparing with simulated blurred images. Inspired by [43], we adopt a Gaussian filter with a standard deviation 2 and window size 7×7 to blur the full clear image repeatedly, thereby obtaining simulated images with different degrees of blur. Then, we use the simulated full blurred images and natural full clear images, which is named as FCSB, to help train our model. Table 4 shows the results. With more degrees of blur in FCSB, the performance is better. However, when $n \geq 4$, the performance has no obvious improvement due to the limited single degree of blur within one image. FCFB is used to train proposed model that achieves the best performance. Since the natural full blurred image contains more degrees of blur within one image, improving the optimization of the model.

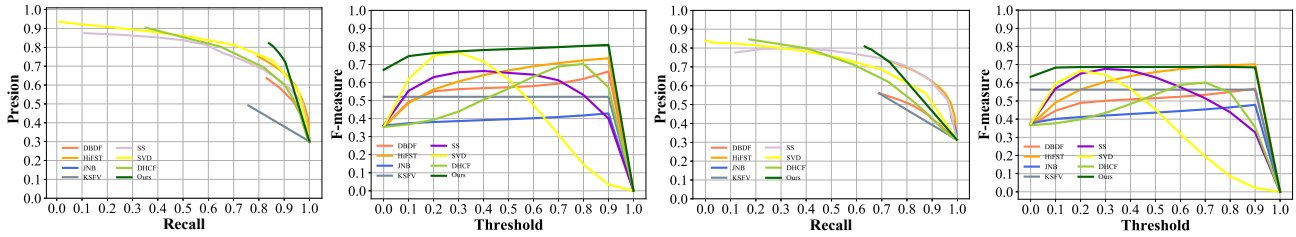
4.3. Comparison with State-of-the-Arts

Thanks to the capability of dual adversarial discriminators that forces the generator to produce an accurate DBD mask, we successfully implement an unsupervised DBD method without using any pixel-level manual annotation. Our method is compared with seven unsupervised methods, including high-frequency multi-scale fusion and sort transform of gradient magnitudes (HiFST) [1], combining deep and hand-crafted features (DHCF) [16], discriminative blur detection features (DBDF) [20], sparse representation and image decomposition (SRID) [21], spectral and spatial



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Figure 5. Visual comparison of DBD masks produced by ours and other unsupervised ones. (a)-(j) are source, ground truth, SVD, DBDF, SRID, KSFV, SS, DHCF, HiFST, and our unsupervised model (without using pixel-level manual annotation). Our model consistently generates DBD masks closest to the ground truth.



(a) Precision-recall curve on CUHK (b) F-measure curve on CUHK (c) Precision-recall curve on DUT (d) F-measure curve on DUT

Figure 6. Comparison of precision-recall curves and F-measure curves of unsupervised DBD methods on CUHK dataset and DUT dataset.

approach (SS) [26], kernel-specific feature vector (KSFV) [15], and singular value decomposition (SVD) [23]. Besides, we compare two full supervised deep learning methods. One is training our generator using cross entropy loss with pixel-level annotation, and the other is the multi-stream bottom-top-bottom network (BTBNet) [43]. In order to implement a fair comparison, we use the recommended parameter settings to generate results, or directly download the results provided by authors.

Quantitative evaluation is shown in Table 3. Our unsupervised model outperforms the second-best HiFST by 9.7% and 2.2% in MAE on CUHK dataset and DUT dataset, respectively, while lowering the MAE significantly. Besides, our unsupervised model achieves competitive performance comparing with supervised BTBNet (*e.g.*, 8.6% and 10.7% performance gaps in F-measure on DUT and CUHK datasets, respectively). Moreover, our unsupervised model is highly efficient with the average testing time of 0.005s. Precision-recall curves and F-measure curves are shown in Figure 6. Our method performs favorably against other methods on both datasets.

Visual comparison results are provided in Figure 5, including various challenging scenes, *i.e.*, cluttered background, low-contrast focused areas, and unfocused foreground. It can be seen that our method highlights focused areas the most accurately (see the last column).

5. Conclusions

We present an effective method to train a deep DBD model without using any pixel-level annotation. The core of our method is the introduction of dual adversarial discriminators, forcing the generator to generate an accurate DBD mask. Thus, the DBD mask can be used to generate a composite clear image and a composite blurred image to simultaneously fool the dual discriminators into believing that the composite images are full clear and full blurred, thereby achieving an implicit manner to define what a defocus blur area is. In addition, we design a bilateral triplet-excavating constraint to assist a balance of the two discriminators, where we encourage the feature-space distance between the composite clear image and another realistic full clear image to get closer, and simultaneously inspire the feature-space distance between the composite blurred image and another realistic full blurred image to be smaller. Extensive experimental results on two widely-used datasets verify that our method outperforms most of unsupervised methods, while owning the fastest calculation speed.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant Nos. 61801077, U1903215, 61872056, 61829102, 61871067, and 61771088.

References

[1] S Alireza Golestaneh and Lina J Karam. Spatially-varying blur detection based on multiscale fused and sorted transform coefficients of gradient magnitudes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5800–5809, 2017. 3, 7

[2] Adam Bielski and Paolo Favaro. Emergence of object segmentation in perturbed generative models. In *Advances in Neural Information Processing Systems*, pages 7256–7266, 2019. 1, 3

[3] Wenchao Du, Hu Chen, and Hongyu Yang. Learning invariant representation for unsupervised image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 4

[5] Shir Gur and Lior Wolf. Single image depth estimation trained via depth from defocus cues. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7683–7692, 2019. 1

[6] Shir Gur, Lior Wolf, Lior Golgher, and Pablo Blinder. Unsupervised microvascular image segmentation using an active contours mimicking neural network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10722–10731, 2019. 3

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1, 2

[8] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *IEEE International Conference on Computer Vision*, pages 9865–9874, 2019. 1

[9] Peng Jiang, Haibin Ling, Jingyi Yu, and Jingliang Peng. Salient region detection by UFO: Uniqueness, focusness and objectness. In *IEEE International Conference on Computer Vision*, pages 1976–1983, 2013. 1

[10] Ali Karaali and Claudio Rosito Jung. Edge-based defocus blur estimation with adaptive scale selection. *IEEE Transactions on Image Processing*, 27(3):1126–1137, 2017. 3

[11] Beomseok Kim, Hyeongseok Son, Seong-Jin Park, Sunghyun Cho, and Seungyong Lee. Defocus and motion blur detection with deep contextual features. In *Computer Graphics Forum*, volume 37, pages 277–288, 2018. 1, 3

[12] Junyong Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee. Deep defocus map estimation using domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12222–12230, 2019. 3

[13] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3623–3632, 2019. 1, 3

[14] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020. 1, 3

[15] Yanwei Pang, Hailong Zhu, Xinyu Li, and Xuelong Li. Classifying discriminative features for blur detection. *IEEE Transactions on Cybernetics*, 46(10):2220–2227, 2015. 3, 7, 8

[16] Jinsun Park, Yu-Wing Tai, Donghyeon Cho, and In So Kweon. A unified approach of multi-scale deep and hand-crafted features for defocus estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1736–1745, 2017. 7

[17] Tal Remez, Jonathan Huang, and Matthew Brown. Learning to segment via cut-and-paste. In *Proceedings of the European Conference on Computer Vision*, pages 37–52, 2018. 1, 3

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015. 5

[19] Parikshit Sakurikar and PJ Narayanan. Composite focus measure for high quality depth maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1614–1622, 2017. 1, 3

[20] Jianping Shi, Li Xu, and Jiaya Jia. Discriminative blur detection features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2965–2972, 2014. 1, 3, 6, 7

[21] Jianping Shi, Li Xu, and Jiaya Jia. Just noticeable defocus blur detection and estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 657–665, 2015. 3, 7

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, pages 1–14, 2014. 5

[23] Bolan Su, Shijian Lu, and Chew Lim Tan. Blurred image region detection and classification. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 1397–1400, 2011. 7, 8

[24] Chang Tang, Xinwang Liu, Xiao Zheng, Wanqing Li, Jian Xiong, Lizhe Wang, Albert Zomaya, and Antonella Longo. DeFusionNET: Defocus blur detection via recurrently fusing and refining discriminative multi-scale deep features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2020. 1

[25] Chang Tang, Xinwang Liu, Xinzhong Zhu, En Zhu, Kun Sun, Pichao Wang, Lizhe Wang, and Albert Zomaya. R^2MRF: Defocus blur detection via recurrently refining multi-scale residual features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12063–12070, 2020. 1, 3

[26] Chang Tang, Jin Wu, Yonghong Hou, Pichao Wang, and Wanqing Li. A spectral and spatial approach of coarse-to-fine blurred image region detection. *IEEE Signal Processing Letters*, 23(11):1652–1656, 2016. 3, 7, 8

[27] Chang Tang, Xinzhong Zhu, Xinwang Liu, Lizhe Wang, and Albert Zomaya. Defusionnet: Defocus blur detection via recurrently fusing and refining multi-scale deep features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2700–2709, 2019. 1, 3

[28] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*, pages 550–558, 2016. 2

[29] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[30] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8198–8207, 2019. 1, 3

[31] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12635–12644, 2020. 1, 3

[32] Guodong Xu, Yuhui Quan, and Hui Ji. Estimating defocus blur via rank of local patches. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5371–5379, 2017. 3

[33] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1162, 2013. 1

[34] Xin Yi and Mark Eramian. LBP-based segmentation of defocus blur. *IEEE Transactions on Image Processing*, 25(4):1626–1638, 2016. 3

[35] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *IEEE International Conference on Computer Vision*, pages 4048–4056, 2017. 1

[36] Dingwen Zhang, Junwei Han, Yu Zhang, and Dong Xu. Synthesizing supervision for learning deep saliency network without human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[37] Ning Zhang and Junchi Yan. Rethinking the defocus blur detection problem and a real-time deep dbd model. In *European Conference on Computer Vision*, pages 617–632. Springer, 2020. 2, 3

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

[39] Wei Zhang and Wai Kuen Cham. Single-image refocusing and defocusing. *IEEE Transactions on Image Processing*, 21(2):873–882, 2012. 1

[40] Zheng Zhang, Yu Liu, Zhihui Xiong, Jing Li, and Maojun Zhang. Focus and blurriness measure using reorganized dct coefficients for an autofocus application. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(1):15–30, 2016. 3

[41] Fan Zhao and Wenda Zhao. Learning specific and general realm feature representations for image fusion. *IEEE Transactions on Multimedia*, 2020. 3

[42] Wenda Zhao, Xueqing Hou, Xiaobing Yu, You He, and Huchuan Lu. Towards weakly-supervised focus region detection via recurrent constraint network. *IEEE Transactions on Image Processing*, 29:1356–1367, 2019. 3

[43] Wenda Zhao, Fan Zhao, Dong Wang, and Huchuan Lu. Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3080–3088, 2018. 1, 3, 6, 7, 8

[44] Wenda Zhao, Fan Zhao, Dong Wang, and Huchuan Lu. Defocus blur detection via multi-stream bottom-top-bottom network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):1884–1897, 2020. 1, 3, 6

[45] Wenda Zhao, Bowen Zheng, Qiuhua Lin, and Huchuan Lu. Enhancing diversity of defocus blur detectors via cross-ensemble network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8905–8913, 2019. 1, 3