MA-GANet: A Multi-Attention Generative Adversarial Network for Defocus Blur Detection

Zeyu Jiang, *Student Member, IEEE,* Xun Xu, *Senior Member, IEEE,* Le Zhang, Chao Zhang, *Member, IEEE,* Chuan Sheng Foo, and Ce Zhu*, *Fellow, IEEE*

***Abstract*—Homogeneous regions and background clutters pose challenges to defocus blur detection. Existing approaches often produce spurious predictions in those regions and relatively low confident predictions in boundary areas. We tackle above issues from two perspectives in this work. Firstly, inspired by the recent success of self-attention mechanism, we introduce channel-wise and spatial-wise attention modules to attentively aggregate features accordingly. Secondly, we propose a generative adversarial training strategy to suppress spurious and low confidence predictions. This is achieved by utilizing a discriminator to identify predicted defocus map from ground-truth ones. In such a way, the defocus network (generator) needs to produce 'realistic' defocus map to minimize discriminator loss. We further show the generative adversarial training allows exploiting additional unlabeled data to improve performance. Moreover, we demonstrate that the existing evaluation metrics for defocus detection often fail to quantify the robustness to thresholding. For more practical comparisons, we introduce a novel AUF_β evaluation metric. Extensive experiments on three public datasets verify the superiority of the proposed methods when compared against the state-of-the-arts approaches.**

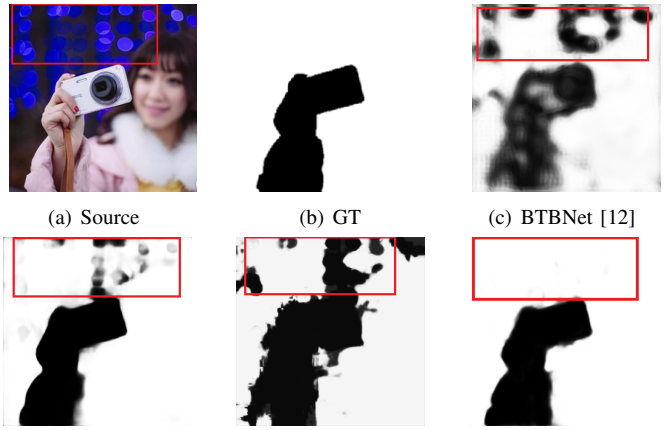
***Index Terms*—defocus blur detection, generative adversarial network, attention module**

I. Introduction

OPTICAL imaging systems produce images with defocus blur when objects are not at the focal region. Defocus blur detection (DBD) aims to separate out-of-focus regions from an image. It has wide applications, including quality assessment [1]–[3], salient object detection [4]–[6], blur magnification [7], image deblurring [8], [9], image refocusing [10], etc.

Traditional defocus blur detection methods usually use the low-level features such as gradient and frequency features [8], [11] to extract the boundaries. Although much progress has been achieved, these hand-crafted methods work well only in limited scenes where the boundary is clear enough to separate in-focus and out-of-focus (blurred) regions. Usually, they fail when trying to separate smoothly blurred regions which do not contain obvious boundary from the smooth in-focus regions, which, are also called homogeneous areas.

To address these issues, deep Convolution Neural Networks (DCNNs) have been applied to defocus blur detection tasks recently. Zhao et al. [12] proposed the BTBNet to integrate the semantic cues and structural information by designing a multi-stream fully convolution network. Tang et al. [13] proposed a network using recurrent fusion and refining modules to integrate multi-level information. Despite the significant progress, these networks do not explicitly consider the interdependencies between features in a channel-wise manner. In addition, they blindly fuse all the detailed features without considering their individual contributions to defocus blur detection. Hence, the results are sometimes interfered by the cluttered background, as shown in Figure 1 (c) and (d) where some low-contrast focal regions are misclassified.



(a) Source (b) GT (c) BTBNet [12] (d) DefuNet [13] (e) LBP [14] (f) Ours

Fig. 1: Cluttered background (rectangular region) have a nonnegligible impact on defocus blur detection. Our method could effectively suppress the interference from background

Recent works on saliency detection and semantic segmentation discovered that the high-level and low-level structural information are usually complementary, where the former captures the global context information and the latter captures the spatial structural details [4], [15], [16]. Both the high-level and low-level information are further integrated for better feature expression by attention mechanism. In the context of defocus detection, Tang et al. [17] employed a channel attention module to select discriminative features by learning the weights of each feature layers. Inspired by these designs, we believe it is beneficial to exploit the separation of high-level and low-level information and introduce attention mechanisms to fuse features for defocus blur detection as well.

In order to improve the discriminative ability of network, we first introduce a channel-wise attention module to explicitly model interdependencies of feature channels by calculating

(Zeyu Jiang and Xun Xu contributed equally to this work.)(Corresponding author: Ce Zhu.)

Xun Xu and Chuan Sheng Foo are with I2R, A-STAR, Singapore.

Zeyu Jiang, Chao Zhang and Ce Zhu are with University of Electronic Science and Technology of China. e-mail: eczhu@uestc.edu.cn. Chao Zhang is also with Sichuan Police College

the correlation of feature maps across channels. Such ability is crucial for accurate detection of low-contrast focal regions and suppressing the interference of background. As shown in [13], [17], the detection maps from high-level features locate an approximate area, while low-level features are good at detecting the sparse and irregular boundaries of defocus regions. Since low-level information alone is prone to the noisy clutters from background and becomes less effective in homogeneous areas, we further propose to use the high-level information to guide the learning of low-level features by providing spatial cues. To this end, we introduce a spatial attention module to guide low-level features with a spatial attention map generated from high-level features. After capturing the desirable high-level information and low-level details, the features are fused together to obtain complementary information and yield final results. Due to the fusion of attentions at multiple levels, the new backbone network is named Multi-Attention Network (MANet). An illustration of MANet is presented in Figure 2.

Apart from the challenge in capturing both high-level and low-level information, existing approaches often produce defocus blur detections with many artifacts which are often with low-confidence, e.g., the blurry and sparse artifacts generated by BTBnet and DefuNet in Figure 1. These artifacts caused visually 'unrealistic' defocus blur detection and this phenomenon is well known as the blurry effect of averaging pixel-wise loss [18], [19]. Because cross-entropy loss defined over individual pixels is adopted, averaging the loss over millions of pixels can be very insensitive to the sparse and irregular wrongly predicted pixels, i.e., the artifacts. To tackle this challenge, we take the approach of Generative Adversarial Network (GAN) [19], [20] to learn a 'high-level' loss function to enforce the prediction to be visually 'realistic'. This is achieved by utilizing a conditional GAN. The generator is implemented by the aforementioned MANet and the discriminator takes both RGB image and defocus prediction map and differentiate the predicted ones from ground-truth ones. Such a generative adversarial training procedure allows discriminator to pick up the 'high-level' difference between prediction and ground-truth, and enforce the generator to produce more 'realistic' prediction. Finally, the discriminator can be seamlessly integrated into the MANet and the we name the whole system as Multi-Attention Generative Adversarial Network (MA-GANet). An interesting observation of MA-GANet is on its ability to fit the real data distribution [20] which is demonstrated as the highly binarized distribution of predicted maps compared against existing ones.

Obtaining labeled images is often expensive since it requires providing pixel-wise annotation while unlabeled images with out-of-focus regions are abundant with almost not cost. The availability of a discriminator network allows us to exploit additional unlabeled data to further improve the quality of generator network. Therefore, we further investigate combining additional unlabeled data during training, a.k.a. semi-supervised learning. In specific, we use additional unlabeled data to enforce the generator to produce more 'realistic looking' defocus map by minimizing the discrminator loss. We observe further improvement in all tasks with additional unlabeled data.

Beyond the novel design of network structure and training strategies, we observe that existing metrics, such as harmonic mean between precision and recall (F_β), Mean Absolute Error (MAE), Intersect over Union (IoU) and Area under the ROC (AUC), are commonly adopted for benchmarking defocus blur detection methods [12], [13], [17], [21]. Among these, we notice that both F_β and IoU require a fixed threshold to binarize the output prediction before calculating metrics. Many works which reported high performance on these two metrics often select the best threshold to maximize the performance. However picking a good threshold without seeing the testing data is non-trivial and it is quite common that there is only a very narrow range of threshold which gives competitive performance. The MAE metric does not require thresholding, but is sensitive to class imbalance. AUC measures the area under true positive rate (TPR) vs. false positive rate (FPR) curve over all possible thresholds. Higher AUC indicates better separation between positive and negative, but it still does not reveal whether there is a large range of threshold values under which the binarized output is accurate. Using one AUC value at a specific threshold (e.g. 0.5) for evaluation may not be fair or representative. To provide a more reasonable measurement of performance, we propose a new evaluation metric based on F_β. We first uniformly sample threshold values from 0 to 1 with a step of 0.01, then the F_β is calculated at each threshold. This will produce a F_β vs. threshold curve, examples are given in Figure 7 (b)(e)(h), and the area under the curve, namely AUF_β, is adopted as a threshold agnostic evaluation metric. A flatter F_β curve and higher mF_β generally indicate a method being more robust to threshold values.

Overall, we summarize our contributions as below.

- A novel Multi-Attention Network (MANet) is proposed to detect defocus regions from images. The end-to-end deep network extracts the interdependencies of features to accurately distinguish defocused blur from homogeneous regions and suppress the interference of background clutter. This has appeared in a preliminary version of this work [22].
- In addition to the contributions made in the preliminary work, we further claim the following contribution in this work. To reduce the artifacts produced by state-of-the-art defocus detection networks, we further propose a generative adversarial training strategy to enforce the output to be 'realistic looking'.
- We further demonstrate that through generative adversarial training, we achieve even higher performance with additional unlabeled data, a.k.a. semi-supervised learning.
- We also propose a new evaluation metric, AUF_β to measure the robustness to threshold values. A higher AUF_β generally indicates the prediction output to be robust to a wider range of threshold.

II. Related Work

Defocus Blur Detection (DBD). Traditional handcrafted based methods mostly focus on the differences of gradients and frequency information of in focus and out-of-focus regions and then extract the edge features because defocus blur usually

blunts object edges. Pang et al. [23] developed a defocus blur detection method based on kernel-specific feature which consists of the information of a blur kernel and the information of an image patch, and the blur regions are distinguished with an SVM. Su et al. [24] found that the first few most significant eigen images of a defocused patch usually have higher weights than in-focus patches. Thus they detected defocus regions by calculating the singular-value of each image pixels. Golestaneh and Karam [25] proposed the method which makes use of the high-frequency DCT coefficients of the gradient magnitudes from multiple resolutions to detect blur regions. Yi and Eramian [14] presented a method which captures the distribution of uniform local binary patterns in blur and non-blur image regions for defocus blur detection. By exploiting the gradient domain information of the corresponding local patches, Xu et al. [26] introduced a ranking-based metric to detect defocus blur regions and they generate the complete blurred regions by a standard propagation method. In order to enhance the discriminative ability for differentiating in-focus and out-of-focus regions, Shi et al. [8] extract a set of defocus blur features including gradient, Fourier domain, and data driven local filter features. These traditional techniques are capable of keeping fine image details. Nevertheless, the hand-crafted features and priors can hardly capture high-level and global semantic knowledge. Therefore, their results can only work well for images with simple structures and are unsatisfying when dealing with complex scenes. Therefore, extracting high-level and enhance the discriminative ability of network is necessary.

Deep CNNs have recently set new standard on a number of visual recognition tasks, including defocus blur detection [12], [13], [21], [27]–[32]. Motivated by such vast successes, Park et al. [29] proposed a unified approach to combine handcrafted and deep blur features at image patch-level and fed them into an fully convolutional network for blurred degree prediction. In [27], two subnetworks are designed to learn global information and local features. Then the probabilities map predicted by the two networks are aggregated and fed into a Markov random field based framework to yield the final prediction map. Based on the observation that defocus blur is sensitive to the image scale, Zhao et al. [12] proposed a multi-stream bottom-top-bottom fully convolutional network to integrate semantic features and detailed information. In order to extract more features, two streams i.e., a forward stream and a backward stream, are used to integrate multi-level features. However, their large number of parameters lead to high storage and computation consumption. Besides, some low-contrast focal areas still cannot be differentiated. Tang et al. [13] proposed a defocus blur detection method based on recurrently fusing and refining of the feature maps. The feature fusion and refinement are performed step-by-step in a cross-layer manner. Zhao et al. [33] introduced a cross-ensemble network to enhance diversity of defocus blur detectors. In [34], A bidirectional residual feature refining network with two branches is constructed to refine the residual features in two directions. The final predicted map is generated by fusing the outputs from two branches. Lee et al. [35] collected a novel dataset with synthetic defocus images for network training and adopted domain adaptation method to address the gap between synthetic images and real ones. Cun and Pun [21] proposed a depth distillation to use depth information for DBD.

Despite the improvement these deep learning based defocus blur detection methods have made, there are still some issues which may make their prediction results unsatisfying. First, most of previous deep learning based DBD methods focus more on acquiring multi-level deep features by building deeper or wider network, without considering the correlations amongst feature maps. Second, existing methods integrate all detailed features without distinction. Thus, their results sometimes are interfered by the background clutter (as shown in Figure 1) and some low-contrast focal region cannot be differentiated.

Attention Mechanism. Attention module has proved its effectiveness in various tasks such as image classification, saliency object detection, video classification, etc. Wang et al. [36] proposed the non-local network mainly exploring effectiveness of non-local operation in spacetime dimension for videos and images. Zhao et al. [4] proposed a pyramid feature attention work for saliency detection. However, their attention modules are less effective in modelling the relationship of feature maps, which is crucial for enhancing the discriminative ability of network. Fu et al. [15] designed two parallel self-attention modules to capture long-range dependencies for semantic segmentation task. In this work, different from existing attention designs, we propose to explicitly model the channel-wise correlation to aggregate features across different channels. Given separated high-level and low-level features, we use the spatial cues of high-level features to weight low-level features beyond trivial fusion.

Structured losses. Defocus blur detection is often formulated as per-pixel classification and each pixel of predicted image is penalized independently from all others because of the adopted per-pixel loss, e.g. cross entropy loss. Structured losses are capable of penalizing the joint configuration of the output to correct high-frequency information. Many methods have considered structured losses, such as the SSIM metric [37], conditional random fields [38], feature matching [39] and losses based on matching co-variance statistics [40]. Recently, some methods [19], [41] applied the Generative Adversarial Networks (GANs) to learn a structured loss for visual data. In [19], a conditional GAN is applied to learn a structured loss to produce more realistic style transfer on images. A generator learns a mapping from random noise and conditional input to the output space and a discriminator learns to discriminate predicted outputs from real ones. In this work, we introduce generative adversarial training to defocus detection to further penalize artifacts in prediction maps. We further demonstrate such design enables using additional unlabeled data to further improve the performance.

III. Methodology

In this section, we first introduce the multi-attention network which learns high-level and low-level features in a two-stream way and fuses both branches to produce defocus prediction. Then we elaborate the generative adversarial training proce-

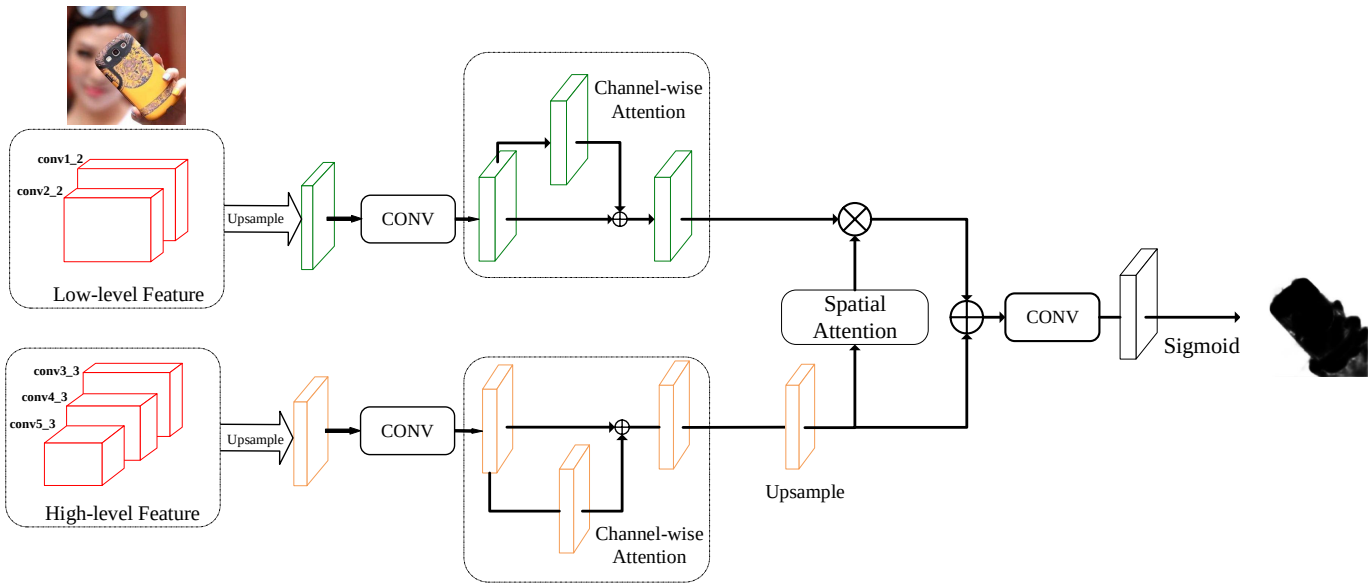


Fig. 2: The framework of the proposed Multi-Attention Network (MANet).

dure which forces the multi-attention network (generator) to produce more 'realistic' predictions.

A. Multi-Attention Network

Accurate detection of low-contrast focal regions and effective suppression of background clutter are main challenges of defocus blur detection. Therefore, it is important to enhance the discriminative ability of network. Most fully convolutional network (FCNs) based defocus blur detection methods do not make full use of the correlations of feature layers, resulting in relatively-low performance in defocus blur detection. Moreover, using low-level features alone could be prone to background clutters and misclassify homogeneous areas. To resolve these issues, we develop an efficient defocus blur detection network taking into consideration the correlation between feature channels and use the spatial attention of high-level features to guide low-level feature learning.

As illustrated in Figure 2 , the pre-trained model VGG-16 [42] is employed as the backbone feature extraction network which produces five basic feature extraction layers denoted by conv1_2, conv2_2, conv3_3, conv4_3 and conv5_3. To transform the original VGG-16 model into a fully convolutional network, we remove the top three fully connected layers of VGG-16. We also delete the five pooling layers to utilize spatial information effectively. The feature extracted by the shallow layers can reflect the fine details which preserve the sharp edges of in-focus objects, while the deep layers could capture high-level spatial extent, which can help avoid in-focus smooth regions being misclassified. Thus we intuitively divide the layers into two groups. Specifically, conv3_3, conv4_3 and conv5_3 are deeper layers. We up-sample the conv4_3 and conv5_3 to the size of conv3_3, then combine all by a cross channel concatenation as the basic high-level features. Meanwhile, shallow layers, conv1_2, conv2_2, are used to exploit detailed information. The similar up-sample operations are carried out to obtain the basic low-level features. Then, both low-level features and high-level features are fed into the channel-wise attention module separately to extract the interdependencies of different feature maps. Afterwards, we use the spatial attention computed from high-level features to guide the learning of low-level features. This step is necessary because high-level features mainly characterize the spatial extent while the low-level features focus on detailed boundaries but are prone to background clutters [43], [44]. The output from both high-level and low-level features are fused together to obtain pixel-level predictions.

1) Channel Attention Module: We employ a channel-wise attention mechanism to emphasize the important features and suppress disturbing information by explicitly modeling channel-wise interdependencies. The module computes responses based on relationships between different channels and improves the representation capability of defocus features.

As illustrated in Figure 3, given a feature $\mathbf{X}^* \in \mathbb{R}^{C\times H\times W}$ we firstly reshape it to $\mathbf{X} \in \mathbb{R}^{C\times N}$, then perform a matrix multiplication between $\mathbf{X}$ and its transpose $\mathbf{X}^T$. Afterwards, the attention map $\mathbf{R} \in \mathbb{R}^{C\times C}$ is obtained by applying a softmax function,

$$r_{ij} = \frac{exp(\mathbf{x}_i \cdot \mathbf{x}_j)}{\sum_{i=1}^{C}(exp(\mathbf{x}_i \cdot \mathbf{x}_j))} \tag{1}$$

where r_{ij} measures the i^{th} channel's influence factor on the j^{th} channel. The more similar two feature maps are, the stronger the correlation will be. Then we apply a matrix multiplication between the transpose of $\mathbf{R}$ and $\mathbf{X}$ to get the output in shape $\mathbb{R}^{C\times H\times W}$. Finally, we multiply a scale factor to the output and add a residual connection to produce the final output $\mathbf{Y}$.

$$\mathbf{Y}_j = \alpha\sum_{i=1}^{C}(r_{ji}\mathbf{X}_i) + \mathbf{X}_j \tag{2}$$

The scale factor α is initialized to zero which means the module have no influence on the input feature maps at the

beginning and gradually learns a proper weight during the training process. It can be inferred from Eq. (2) that the resultant feature map $\mathbf{Y}$ is a weighted sum of all channels and the original map. Therefore, it models the interdependencies across feature channels. The similar feature maps achieve mutual gains, thus emphasizing desired features, gaining better representation of defocus features and enhancing the discriminative ability. In order to make full use of feature correlations, channel-wise attention module is employed to both high-level and low-level features.

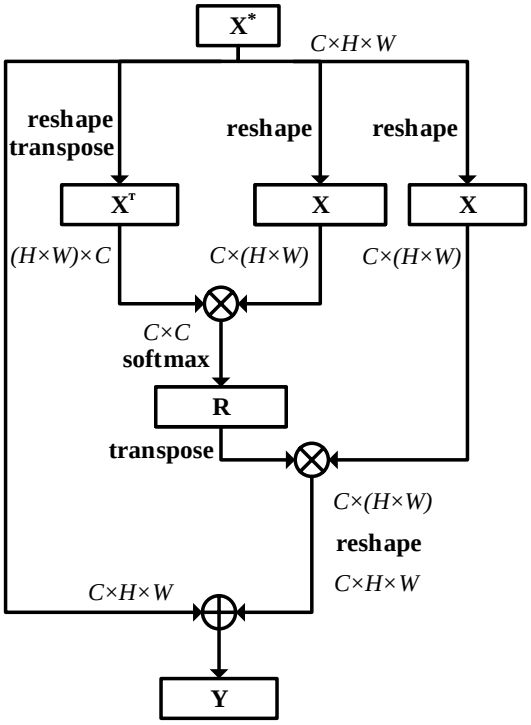


Fig. 3: The details of channel attention module."⊕" and "⊗" denote matrix multiplication and element-wise summation, respectively.

2) Spatial attention module: The low-level cues are essential to defocus blur detection to help refine the sparse and irregular detection regions. By utilizing deep CNNs, we could extract fine detailed information. However, most existing defocus blur detection methods integrate all features without distinction, which leads to information redundancy. More importantly, some detailed information would lead to a performance degradation. For instance, some out-of-focus regions with strong detailed information may be mistakenly regarded as in-focus regions as in Figure 1. To address this issue, we propose a spatial attention module to adaptively emphasize desired low-level features. As illustrated in Figure 2, the outputs of low-level channel-wise attention module will be fed into a spatial attention module which utilizes the high-level spatial cues to adaptively emphasize low-level details. Specifically, $\mathbf{X}^h \in \mathbb{R}^{C\times H\times W}$ stands for high-level features and $\mathbf{X}^l \in \mathbb{R}^{C\times H\times W}$ stands for low-level features. In order to increase receptive field without additional computation cost, two consecutive atrous convolutions are applied to extract

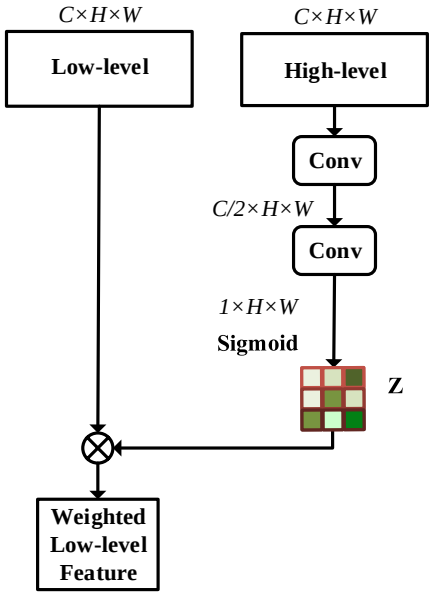


Fig. 4: The details of spatial attention module. "⊗" denotes element-wise product

spatial information (see Figure 4). After mapping the extracted features to [0,1] by a sigmoid function, we obtain the final attention weight map $\mathbf{Z}$. The final output of the low-level features $\tilde{\mathbf{X}}^l$ is acquired by weighting low-level feature $\mathbf{X}^l$ with spatial attention weight map $\mathbf{Z}$ as,

$$\tilde{\mathbf{X}}^l = \mathbf{X}^l \circ \mathbf{Z} \tag{3}$$

where $\circ$ indicates element-wise product. Since the high-level features capture the spatial extent well, such re-weighting could help suppress the erroneous prediction in background clutter.

3) Loss Function: Cross entropy loss is widely adopted by existing methods [33], [34] for training defocus detection network. However, as seen from Figure 5, the two classes (in-focus and out-of-focus regions) are often highly imbalanced with out-of-focus region being the majority. Cross entropy loss is calculated over each individual pixel and the averaged loss is used for training, as a result it is very sensitive to the imbalance and will bias the network towards predicting out-of-focus.

To tackle this issue, we introduce the intersection-over-union (IoU) loss. IoU is a commonly used evaluation criterion for the segmentation problem. Given an image and its corresponding label, IoUs give the similarity between the predicted region and ground truth region and calculate the area of intersection divided by the union area of the two regions. The IoU measure can effectively take into consideration the imbalance problem. For example, if a trivial solution predicts every pixel to be out-of-focus region, the intersection between the predicted region and ground-truth and IoU metric would both be zero. Therefore, it potentially improves the defocus detection by penalizing small IoU metric. Luckily, this objective can be easily converted into the IoU loss,

$$\mathcal{L}_{IoU} = 1 - \frac{|f(\mathbf{X}) \circ \mathbf{Y}|}{|f(\mathbf{X})| + |\mathbf{Y}| - |f(\mathbf{X}) \circ \mathbf{Y}|} \tag{4}$$

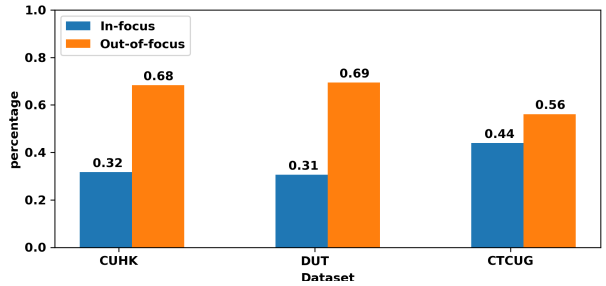


Fig. 5: Distributions of in-focus and out-of-focus pixels in different datasets.

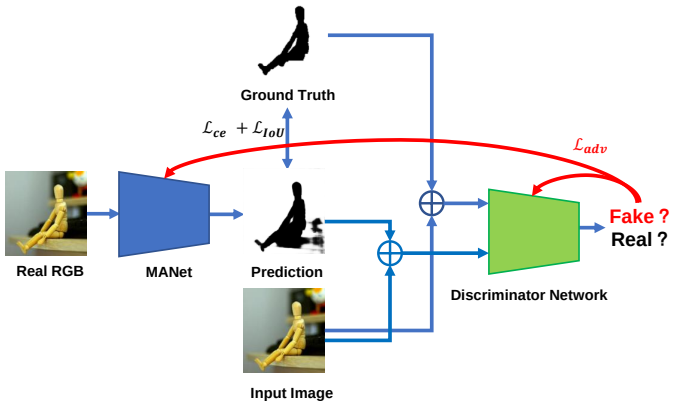


Fig. 6: Framework of the proposed Multi-Attention Generative Adversarial Network (MA-GANet). "⊕" denotes concatenation operation across channels.

where $f(\mathbf{X})$ is the probabilistic output of MANet and $|\cdot|$ is the L1 norm.

Despite being robust to class imbalance, IoU may result in diminishing gradient when there is no overlap between prediction and ground-truth, i.e. $|f(\mathbf{X}) \circ \mathbf{Y}|$. To overcome this challenge, we also employ cross entropy loss to complement IoU loss. The pixel-wise cross entropy Loss between prediction $f(\mathbf{X})$ and the ground truth mask $\mathbf{Y}$ is calculated as:

$$\mathcal{L}_{CE} = |\mathbf{Y} \circ \log f(\mathbf{X}) + (\mathbf{1} - \mathbf{Y}) \circ \log(\mathbf{1} - f(\mathbf{X}))| \quad (5)$$

The combined loss function to optimize is shown in Eq. (6)

$$\lambda_{CE}\mathcal{L}_{CE} + \lambda_{IoU}\mathcal{L}_{IoU} \quad (6)$$

Where λ_{CE} and λ_{IoU} represents the weight of cross entropy loss and IoU loss respectively.

B. Discriminative Network

Defocus blur detection is formulated as a per-pixel classification problem as above. The known issues with such pixel-wise loss result in artifacts in output prediction. Thus, we propose to adopt a conditional Generative Adversarial Network (cGAN). The discriminator learns to differentiate predicted defocus map from ground-truth ones, serving as a 'structural loss' to penalize out-of-distribution predictions. It allows prediction network, also known as the generator, to produce more 'realistic' looking defocus outputs.

1) Conditional Generative Adversarial Networks: A Generative Adversarial Network (GAN) [20] consists of two adversarial networks: a generator G, and a discriminator D which are trained in a min-max game manner. The generator G is trained to map an input, e.g., a random noise vector, to an output space which is similar to the data distribution. Meanwhile, the discriminator is optimized to distinguish synthesized data from the true data distribution. The conditional variant of GAN (cGAN) [19] further takes a data sample as input, in contrast to a random noise vector, so that the generator could be realized as arbitrary network transforming data sample in one domain into another domain. More formally, we define the generator as a mapping from input RGB image to defocus map, $G : \{\mathbf{X} \rightarrow \mathbf{Y}\}$. Such a generator can be instantiated as the MANet introduced in previous sections. A discriminator is also defined as a mapping from a pair of RGB image and the associated defocus map, $D : \{\{\mathbf{X}, \mathbf{Y} \rightarrow r\}\}$ where $r \in [0, 1]$ is a probabilistic prediction. The loss function cGAN is thus written as,

$$\begin{aligned}\mathcal{L}_{cGAN} =& \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\log D(\mathbf{X}, \mathbf{Y})\right] + \\ & \mathbb{E}_{\mathbf{X}}\left[\log(1 - D(\mathbf{X}, G(\mathbf{X})))\right]\end{aligned} \quad (7)$$

We follow the practice in [19] to omit the noise vector from the generator as a diverse output can be achieved by the dropout layers in the MANet.

As we discussed above, by optimizing the min-max objective, the discriminator is able to help the generator produce predictions which are closer to the true data distribution. The defocus detection network, known as MANet, takes RGB image as input and yields prediction maps as output. It is natural to take the MANet as the generator to realize the mapping from input image to defocus prediction.

To construct the discriminator, we are aware that the low-frequency components are already constrained by the pixel-wise loss, e.g., the cross entropy loss and IoU loss. Importantly, the high-frequency components, which are accountable for the commonly observed artifacts, are not easily captured by the existing loss functions. Thus, the discriminator only needs to focus on the correctness in high-frequency components. Due to many spatially repetitive patterns in the image, we adopt PatchGAN [19] as the discriminator. It consists of K fully convolution layers, each pixel of the $M \times M$ output feature map is classified into 'real' or 'fake'. Each pixel of feature map determines if the corresponding rectangle patch in the input domain looks 'real' or not, thus termed as PatchGAN by [19].

It has been widely observed that training GAN is prone to instability, when the target distribution differs too much from model distribution, the learning gradient for generator may vanish or explode [45]. To tackle this issue, the Spectral Normalization technique was introduced to stabilize the training of GAN so that the learning gradient is well normalized [46]. In specific, for each convolution kernel $\mathbf{W}$ in PatchGAN discriminator, we apply the following normalization,

$$\hat{\mathbf{W}} = \frac{\mathbf{W}}{\sigma(\mathbf{W})}, \quad \sigma(\mathbf{W}) = \max_{\mathbf{h}:\mathbf{h}\neq\mathbf{0}} \frac{||\mathbf{W}\mathbf{h}||_2}{||\mathbf{h}||_2} \quad (8)$$

where $\sigma(\mathbf{W})$ is the largest singular value of $\mathbf{W}$.

We integrate the MANet, as generator, and the discriminative network into a unified framework, termed as Multi-Attention Generative Adversarial Network (MA-GANet) with illustration in Figure 6. The final training loss function combines both regular defocus training loss, i.e., pixel-wise cross-entropy loss $\mathcal{L}_{CE}$, IoU loss $\mathcal{L}_{IoU}$, and the generative adversarial training loss $\mathcal{L}_{cGAN}$. The training objective writes as in Eq. (9).

$$\min_{G}\{\lambda_{CE}\mathcal{L}_{CE} + \lambda_{IoU}\mathcal{L}_{IoU} + \lambda_{cGAN}\max_{D}\mathcal{L}_{cGAN}(G, D)\} \tag{9}$$

C. Semi-Supervised Learning

Training defocus detection is subject to the high cost of acquiring labeled data. Annotating defocus map is particularly challenging as labels for every pixel must be provided. We investigate in this section using additional unlabeled data to further improve the performance. Given a new batch of data with half being labeled $\{\mathbf{X}_i^l, \mathbf{Y}_i^l\}$ and others being unlabeled $\{\mathbf{X}_j^u\}$, we take the following steps to update discriminator and generator. For discriminator, we still update with labeled data only by minimizing the cGAN loss in Eq. (7). For updating discriminator, we take a two-step approach. In the first step, we compute the combined loss $\mathcal{L}_l^* = \lambda_{CE}\mathcal{L}_{CE} + \lambda_{IoU}\mathcal{L}_{IoU} + \lambda_{cGAN}\mathcal{L}_{cGAN}(G, \hat{D})$ with labeled data where $\hat{D}$ is the fixed discriminator network. In the second step, we compute discrminator loss $\mathcal{L}_u^* = \lambda_{cGAN}\mathcal{L}_{cGAN}(G, \hat{D})$ with unlabeled data only. Then we accumulate the gradient computed from these two steps $0.5 * \nabla_{\Theta_G}\mathcal{L}_l^* + 0.5 * \nabla_{\Theta_G}\mathcal{L}_u^*$ as the final gradient for updating generator where Θ_G and Θ_D are parameters for generator and discriminator respectively.

D. Training and Inference Details

In this section, we elaborate the details of training MA-GANet. We initialize backbone's parameters from a VGG-16 network pretrained on the ImageNet and initialize discriminator's parameters randomly from a Gaussian distribution with mean 0 and standard deviation 0.02. Since it involves solving a min-max game, we follow a two-step iterative optimization procedure. In the first step we sample a mini-batch $\{\mathbf{X}_i, \mathbf{Y}_i\}$ and do forward pass on MANet (generator) with output $f(\mathbf{X})$. Then we concatenate input image and ground-truth/prediction as $\{(\mathbf{X}_i, \mathbf{Y}_i)_i\}$ and $\{(\mathbf{X}_i, f(\mathbf{X}_i))_i\}$, respectively. The discriminator is updated by the calculated adversarial loss. In the second step, we recalculate the adversarial loss with fixed discriminator, and the total loss is minimized by one step gradient descent. The overall training algorithm is presented in Algo. 1. During the inference stage, only the generator is employed to map input RGB image to defocus prediction. When additional unlabeled data is available, we first implement a fixed steps of warm-up, i,e, training the model in a supervised manner. Afterwards, we turn on semi-supervised training following Sect. III-C.

Algorithm 1: Training MA-GANet

Input: Training Images $\mathcal{X}$, labels $\mathcal{Y}$
Output: Network parameters $\Theta = \{\Theta_D, \Theta_G\}$
Initialize the network.
for *number of training iterations* **do**
  # Train discriminator $D(\mathbf{X}, \mathbf{Y}; \Theta_D)$
  Calculate L_{cGAN} by Eq. (7)
  Update discriminator by maximizing L_{cGAN}
  # Train generator $G(\mathbf{X}, \mathbf{Y}; \Theta_G)$
  Freeze the parameters of discriminator Θ_D
  Calculate $\mathcal{L}_{CE}$, $\mathcal{L}_{IoU}$ and recalculate $\mathcal{L}_{cGAN}$
  Update the MANet by minimizing Eq.(9)
  Unfreeze the parameters of discriminator Θ_D

IV. Experiment

A. Datasets

In our experiments, we demonstrate on three publicly available datasets with pixel-level annotations. **Shi et al.'s dataset (Shi's Dataset)** [47] consists of 704 partially defocus blurred images. We divided 704 defocus blur images with pixel-level masks into two parts, i.e., the first 604 images for training and the rest 100 images for testing as [13]. **DUT** [12] is a defocus blur detection dataset proposed by Zhao et al. which consists of 600 training images and 500 test images with pixel-level annotations. It is a more challenging dataset because images have multi-scale focused areas, i.e. low contrast focal regions and strong background clutter. **CTCUG** [17] is a newly collected test dataset which contains 150 images with manual pixel-wise annotations. All 150 images are used for testing only in this work and training is carried out with DUT dataset. Several challenging cases are considered in the CTCUG, such as complex background, in-focus areas with low contrast, in-focus and out-of-focus foreground, and same class of objects with different defocus condition. For semi-supervised learning, we use additional unlabeled data from **FRD dataset** [48] which was originally proposed for weakly supervised defocus detection with 5000 images each annotated with a box indicating the rough area of in-focus region. We only use these images as unlabeled data for training.

B. Evaluation Metric

We first consider several widely adopted single numeric evaluation metrics, including F_β, Mean Absolute Error (MAE), Intersect over Union (IoU) and ROC area under the curve (AUC). F_β measures the harmonic mean between precision and recall. MAE measures the absolute difference between prediction and ground-truth with both normalized to between 0 and 1. IoU measures the overlap between binarized prediction and ground-truth. AUC measures defocus detection as binary classification problem. All except MAE are higher the better.

To further provide insight into the behavior of all competing methods, we evaluate precision-recall curve, F_β curve and ROC curve. The F_β curve plots the F_β vs. threshold where F_β is the harmonic mean of precision and recall with defocus

prediction map binarized at each threshold. A good performing method should not only achieve higher point in this curve also maintain high F_β over all thresholds. Therefore, based on the F_β curve, we further propose a new numeric evaluation metric AUF_β which is the area under the F_β curve. It simultaneously measures the accuracy and robustness of a method.

C. Implementation Details

We uses the VGG-16 [42] as backbone network for feature extraction. All existing datasets have limited number of training samples which hampers generalization trained deep neural network. In order to improve generalization and reduce overfitting, we apply data augmentation to each origin image by sequentially random horizontal flipping, random rotation between -0.15π and 0.15π, resizing and cropping. Specifically, resizing refers to rescaling the image to 384×384 pixels and cropping further crops a 320×320 patch from resized image. The whole network is optimized by Adam optimizer [49] and The learning rate is initialized to $4e-4$. The momentum is 0.9 and the weight decay is $5e-4$. The training batch size is 8. We train 1,000 epochs on the all datasets. For semi-supervised learning, we warm-up for 200 epochs with the same setting as above. Afterwards, we following the setting described in Sect. III-C with 4 labeled samples and 4 unlabeled samples in a minibatch.

D. Comparison with the state-of-the-art methods

1) Competing Methods: We compare our method with other 11 state-of-the-art approaches, including analyzing spatially-varying blur (ASVB) [50], discriminative blur detection features (DBDF) [8], spectral and spatial approach (SS) [51], local binary patterns (LBP) [14], high-frequency multiscale fusion and sort transform (HiFST) [25], bottom-top-bottom network (BTBNet) [12], defocus map estimation using domain adaptation (DMENet) [35], depth Distillation (DD) [21], recurrently fusing and refining multi-scale deep features (DefuNet) [13] and its extended version [17] denoted as DefuNetV2. For a fair comparison, all the deep-learning based methods take pretrained VGGNet [42] as backbone network.

For our own methods, we refer MANet to the vanilla model without generative adversarial training proposed in Sect. III-A. We further evaluate MA-GANet which incorporates generative adversarial training. Finally, we evaluate MA-GANet with semi-supervised training (MA-GANet-s).

2) Quantitative Comparison: We first provide numerical comparisons against state-of-the-art methods in Table I. All five metrics are reported on three datasets. Due to inconsistent data splits are adopted by BTBNet, CENet and DD for Shi's dataset, we defer the comparison to these three methods to Table III. From both tables, we make the following observations. First, IoU and AUF_β are more indicative than the other three metrics. For example, on Shi's dataset, the absolute gaps between MA-GANet (best) and ASVB (worst) are 0.222 for F_β, 0.628 for MAE, 0.882 for IoU, 0.360 for AUC and 0.754 for AUF_β, respectively. Given all metrics normalized to between 0 and 1, more significant gaps are observed for IoU and AUF_β. This suggests future comparisons should focus more on these two metrics. In addition, we make clear observation that our MANet already achieves higher performance than DefuNetV2 on most metrics. This is attributed to the advantage of spatial attention module. The MA-GANet with generative adversarial training achieves the state-of-the-art performance under standard fully supervised training settings on all three datasets, demonstrating the effectiveness of discriminator network. Finally, with additional unlabeled MA-GANet-s further improves performance suggesting the effectiveness of exploiting unlabeled data.

We also observe that the CTCUG dataset is the most challenging one with all methods achieving performance consistently lower than DUT and Shi's datasets, suggesting future works should pay more attention to transfer learning benchmarking on this dataset. Finally, under Zhao's split [12], [33] in Table III, our MANet and MA-GANet still outperform BTBNet, CENet and DD with a large margin, which is consistent with observations made from alternative data split adopted in Table I.

We further present the three types of curves, precision and recall (PR) curve, F_β curve and receiver operating curve (ROC), to compare all methods in Figure 7. First, we observe a clear margin between MANet (pink) and other state-of-the-art methods. This is supported by the high profile of all three curves for MANet. Combined with generative adversarial training, the final MA-GANet clearly beats all existing models. Moreover, the F_β curve further reveals the robustness to the choice of threshold. Though some methods report very high F_β measure, it is observed from the figure that these methods are highly sensitive to the choice of threshold. In contrast, our proposed MA-GANet yields robust defocus prediction performance within a large range of threshold.

Running Efficiency In addition to improved results, our method is also efficient in inference. We run inference on a workstation with an Intel 3.4GHz CPU with 32 GB memory and a single Nvidia GTX Titan Xp. The average inference time for an image of different methods are shown in Table II. The non-deep learning approaches run inference on CPU and significantly slower than deep learning approaches running on GPU. MA-GANet is fast on all three datasets, achieving 15-20 fps in average.

3) Qualitative Comparisons: We present qualitative examples of defocus prediction for all competing methods in Figure 8. We first observe that all non-deep learning approaches produce noisy (e.g. LBP)and ambiguous (e.g. SS) predictions. These observations suggest the hand-crafted methods are prone to unclear boundaries and homogeneous areas and thus fail to produce clear defocus detection. We further observe that the deep learning based methods (e.g. CENet, DefuNet, DefuNetV2 and our proposed methods) produce significantly better defocus detection than the non-deep counterparts. Nevertheless, due to the absence of considering interdependencies between spatial and feature channels, the existing approaches often fail on images with cluttered background and low-contrast areas.

In specific, the state-of-the-art DefuNetV2 failed to separate the near focus person. For images with cluttered background, image 2, 4 and 8, DefuNetV2 misclassified in-focus and

out-of-focus regions. In contrast, with the ability to capture feature interdependencies and fusion of low-level and high-level features, our MANet produces much clearer defocus predictions. More interestingly, by introducing the discriminator, the final MA-GANet further removes many artifacts produced by MANet and produces more cleaner and binarized defocus outputs. With additional unlabeled data, the semi-supervised approach (MA-GANet-s) produces results most similar to the ground-truth (GT).

E. Ablation Study

In this section, we evaluate the effectiveness of the two components proposed with both quantitative and qualitative ablation studies.

1) Effectiveness of Attention Modules: We extensively analyze the impact of attention modules. Specifically, we first decompose the attention modules into high-level feature branch (HL) where only conv3_3, conv4_3 and conv5_3 are used; low-level feature branch (LL) where conv1_2, conv2_2 are used; high-level channel-wise attention module (HCA) where the channel-wise attention module is applied to high-level branch; low-level feature channel-wise attention module (LCA); and spatial attention module (SA) where spatial attention module is applied. We choose the VGG model with HL alone as the baseline model. Then, we gradually add low-level feature branch (LL), high-level channel attention module (HCA), low-level channel attention module (LCA) and spatial attention module (SA) to augment the baseline model. The combination of IoU loss and CE loss is adopted for all ablative models. We carried out experiments with all ablative models on DUT dataset since it is the most representative dataset for defocus detection. All five metrics are evaluated for each ablative model with results reported in Table IV. We observe from these results that, first, all components are contributing positively with consistent improvements in all metrics. In particular, fusing HL with LL is important with more than 0.03 improvement in MAE from baseline. We also notice it is important to apply spatial attention to guide the low-level feature learning from high-level branch, the final model observes 0.01 improvement in MAE from with SA module and either with HCA or LCA alone SA does improve consistently. The above observations suggest the importance of fusing low-level and high-level features via spatial attention module.

2) Advantage of Generative Adversarial Network: We introduce the discriminative network to learn a proper structured loss to encourage 'realistic' defocus predictions. Here we investigate the impact of adversarial loss. Specifically, we compare MANet and MA-GANet with results reported on DUT dataset in Table IV. It can be observed that after adding the adversarial loss to the original MANet, the performance of the network has been further improved on 4 out of the 5 evaluation metrics. More specifically, large improvements are logged for MAE, over 10% relatively, and AUF_β both of which are more indicative than others. These observations suggest the efficacy of proposed generative adversarial training mechanism.

3) Semi-Supervised Learning: : We investigate the effectiveness of exploit additional unlabeled data to further improve the defocus detection performance. We compare the performance on DUT dataset with additional unlabeled data (MA-GANet+SSL) in Table IV. We observe all five metrics are further improved with these additional unlabeled data.

4) Qualitative Ablation Study: To visually further demonstrate the impact of attention modules and generative adversarial training we compare baseline model (Baseline), HL+LL+SA (Baseline+SA), HL+LL+HCA+LCA (Baseline+CA), MANet, MA-GANet and MA-GANet+SSL (MA-GANet-s). Qualitative results are shown in Figure 9.

Spatial Attention Effectiveness We first investigate the effect of spatial attention module by comparing the defocus predictions by Baseline and MANet columns. The networks without spatial attention are unable to adaptively select correct spatial extent, hence, its detection results are influenced by background clutters. For example, the prediction in the second row by Baseline is interfered by the background and have unclear boundaries around the rose, whereas the Baseline+SA is capable of producing sharper and better results. Besides, the smooth out-of-focus areas on macarons (third row) are mistakenly predicted as in-focus region without spatial attention module (Baseline and Baseline+CA) while the MANet (with spatial attention) could effectively distinguish the out-of-focus regions.

Channel Attention Effectiveness We compare Baseline+SA against MANet to highlight the importance of channel attention module. Apparently, compared with outputs from MANet, some low contrast in-focus regions by Baseline+SA are misclassified and the cluttered background interferes the detection result. This is because the network without channel attention is unable to extract the inter-dependencies of features thus hindering the discriminative ability. To be specific, in the first row, almost all of the clear regions are wrongly taken as blurred ones by Baseline+CA. Even with the help of spatial attention module, there is still half of the in-focus pixels being mispredicted because the network cannot effectively extract the high-level information.

Generative Adversarial Training Finally, we investigate the effect of introducing generative adversarial training. It can be seen that with the help of generative adversarial networks, MA-GANet yields better detection results with sharper boundaries and finer details. More specifically, some smooth in-focus regions, i.e., part of the arm in the fourth row, are misclassified by MANet, while MA-GANet could give an accurate prediction and preserve the boundary information of the in-focus objects well.

V. Additional Study

In this section, we investigate the impact of hyperparameters that could affect the performance of the proposed network.

A. Loss Weight

In Sec.III, we proposed multiple loss functions including cross entropy loss, IoU loss and adversarial loss. We study the importance of each loss by adjusting the weights, λ_{CE},

TABLE I: Quantative comparison of F-measure, MAE, IoU, AUC, and mean F_β (AUF_β) scores. The best two results are shown in red and blue colors, respectively.

Datasets	Metric	ASVB	DBDF	SS	LBP	HiFST	BTBNet	CENet	DefuNet	DMENet	DD	DefuNetV2	MANet	MA-GANet	MA-GANet-s
Shi's	$F_\beta\uparrow$	0.731	0.841	0.787	0.866	0.865	—	—	0.917	0.914	—	0.925	0.951	0.953	0.955
	MAE$\downarrow$	0.636	0.323	0.298	0.186	0.232	—	—	0.116	0.155	—	0.102	0.096	0.084	0.080
	IoU$\uparrow$	0.04	0.547	0.742	0.757	0.732	—	—	0.833	0.826	—	0.845	0.869	0.886	0.882
	AUC$\uparrow$	0.592	0.839	0.829	0.873	0.829	—	—	0.922	0.880	—	0.924	0.951	0.952	0.953
	$AUF_\beta\uparrow$	0.192	0.761	0.800	0.834	0.828	—	—	0.920	0.702	—	0.930	0.938	0.946	0.948
DUT	$F_\beta\uparrow$	0.747	0.802	0.784	0.874	0.866	0.887	0.903	0.922	0.930	0.932	0.952	0.950	0.954	0.958
	MAE$\downarrow$	0.651	0.369	0.296	0.173	0.302	0.190	0.135	0.115	0.314	0.113	0.082	0.078	0.070	0.068
	IoU$\uparrow$	0.052	0.529	0.529	0.782	0.635	0.803	0.839	0.857	0.846	0.857	0.887	0.899	0.907	0.908
	AUC$\uparrow$	0.582	0.779	0.779	0.859	0.866	0.856	0.898	0.889	0.846	0.901	0.952	0.969	0.967	0.974
	$AUF_\beta\uparrow$	0.219	0.730	0.730	0.840	0.792	0.844	0.881	0.894	0.753	0.898	0.952	0.934	0.952	0.954
CTCUG	$F_\beta\uparrow$	0.605	0.740	0.741	0.805	0.785	0.827	—	0.891	0.845	0.905	0.899	0.899	0.910	0.926
	MAE$\downarrow$	0.505	0.344	0.302	0.242	0.267	0.177	—	0.138	0.301	0.122	0.127	0.119	0.105	0.092
	IoU$\uparrow$	0.066	0.466	0.670	0.650	0.606	0.731	—	0.758	0.754	0.774	0.761	0.788	0.802	0.809
	AUC$\uparrow$	0.649	0.770	0.827	0.821	0.839	0.910	—	0.897	0.899	0.926	0.930	0.938	0.942	0.957
	$AUF_\beta\uparrow$	0.266	0.637	0.731	0.681	0.754	0.832	—	0.865	0.693	0.869	0.866	0.892	0.900	0.912

TABLE II: Average running time(s) for an image of different methods on different datasets

	Methods	ASVB	DBDF	SS	LBP	HiFST	BTBNet	CENet	DefuNet	DefuNetV2	MA-GANet
	CPU/GPU	CPU	CPU	CPU	CPU	CPU	GPU	GPU	GPU	GPU	GPU
Datasets	Shi's	2.04	214.83	2.76	57.34	2576.24	25	0.064	0.094	0.097	**0.062**
	DUT	1.59	110.37	1.20	30.38	1169.57	25	0.064	0.056	0.059	**0.050**
	CTCUG	1.65	120.24	2.01	34.55	1204.71	—	—	—	0.068	**0.060**

TABLE III: Comparison on the datasplit proposed by [12] for Shi's dataset

Methods	MAE $\downarrow$	$F_\beta\uparrow$	IoU $\uparrow$	AUC$\uparrow$	$AUF_\beta\uparrow$
BTBNet	0.109	0.950	0.909	0.980	0.913
CENet	0.060	0.956	0.921	0.951	0.944
DD	0.048	0.966	0.936	0.972	0.960
MANet	0.054	0.960	0.926	0.986	0.949
MA-GANet	0.043	0.969	0.940	0.983	**0.968**
MA-GANet-s	**0.040**	**0.970**	**0.941**	**0.986**	**0.968**

TABLE IV: Ablation analysis of the different components combinations.

Methods	MAE	F_β	IoU	AUC	AUF_β
Baseline	0.132	0.924	0.865	0.947	0.914
HL+LL	0.101	0.937	0.886	0.951	0.932
HL+LL+HCA	0.092	0.938	0.878	0.955	0.926
HL+LL+HCA+SA	0.088	0.944	0.888	0.962	0.932
HL+LL+HCA+LCA	0.089	0.938	0.883	0.965	0.928
HL+LL+SA	0.087	0.940	0.882	0.963	0.929
MANet	0.078	0.950	0.899	0.969	0.934
MANet+adv (MA-GANet)	0.070	0.954	0.907	0.967	0.952
MA-GANet+SSL (MA-GANet-s)	**0.068**	**0.958**	**0.908**	**0.974**	**0.954**

λ_{IoU} and λ_{cGAN} for each loss, respectively. The results are shown in Table V. We observe a combination of cross entropy loss and IoU loss consistently improves the performance, due to the ability to handle imbalanced samples. Moreover, by introducing a discriminator moderately (0.001), we achieve the best performance, particularly on MAE and AUF_β metrics.

TABLE V: The impact of different loss function.

λ_{CE}	λ_{IoU}	λ_{cGAN}	MAE$\downarrow$	$F_\beta\uparrow$	IoU$\uparrow$	AUC$\uparrow$	$AUF_\beta\uparrow$
1.0	0	0	0.078	0.950	0.899	**0.969**	0.934
1.0	1.0	0	0.076	0.950	0.902	0.960	0.938
1.0	1.0	0.001	**0.070**	**0.954**	0.907	0.967	**0.952**
1.0	1.0	0.01	0.077	0.951	0.901	0.969	0.949
1.0	1.0	0.1	0.088	0.948	0.879	0.959	0.927
1.0	0	0.001	0.072	0.953	0.905	0.967	0.947
1.0	5.0	0.001	0.075	0.951	**0.909**	0.966	0.947

B. Discriminator Receptive Field

We evaluate the impact of varying the patch size N of our discriminator receptive fields by adjusting the depth of GAN discriminator and Table VI quantifies the effects using several evaluation criterion. Note that elsewhere in this paper, unless specified, all experiments use 70×70 PatchGANs, and for this section all experiments use an CE+IoU+cGAN loss. Applying a 70×70 PatchGAN is capable of producing promising results and yielding better performance. Scaling beyond this to the full 256 × 256 PatchGAN produces inferior results. This may be explained as that 256 × 256 PatchGAN has much more parameters than the 70×70 PatchGAN redenring training more difficult.

C. Distribution Modelling

It is known that GAN fits the true data distribution [20]. In this section, we demonstrate through comparing the frequency of pixel-wise defocus prediction between MANet and MA-GANet in Figure 10. The x-axis is the probabilistic defocus prediction and y-axis is the average number of pixels per

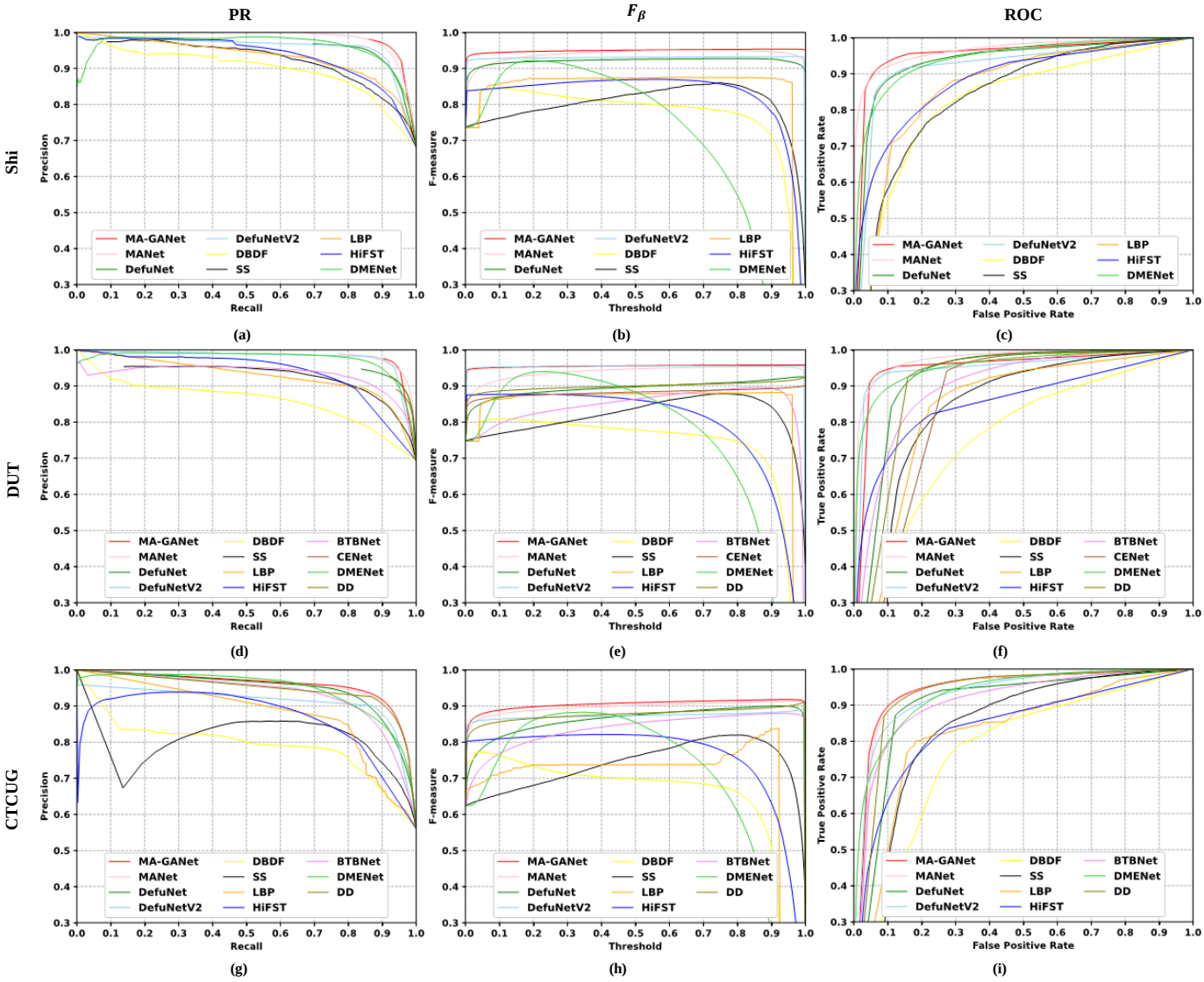


Fig. 7: Comparison of the PR curves, F_β curves and ROC curves of the different methods on datasets.

TABLE VI: The impact of different discriminator receptive field

Receptive field	MAE	F_β	IoU	AUC	AUF_β
1×1	0.084	0.948	0.878	0.957	0.939
16×16	0.073	0.954	0.901	0.964	0.950
70×70	**0.070**	**0.954**	**0.907**	**0.967**	**0.952**
256×256	0.088	0.941	0.868	0.961	0.937

image. It is clear that MA-GANet generates more binarized defocus prediction than MANet, suggesting incorporating generative adversarial training helps fit the ground-truth data distribution.

VI. Conclusion

We propose a novel method named Multi-Attention Network (MANet) for accurate and efficient defocus blur detection. Specifically, a channel-wise attention module is employed to both low-level features and high-level features for better feature representation. A spatial module is applied to the low-level features, so as to focus more on desired details and suppress the background clutter. In addition, we combine MANet, as generator, with a PatchGAN discriminator into a Multi-Attention Generative Adversarial Network (MA-GANet) which produces high fidelity defocus prediction. Finally, we exploit additional unlabeled with MA-GANet to further improve the defocus detection quality. Extensive experimental results demonstrate our method outperforms the state-of-the-art methods in terms of both existing evaluation metrics and our newly proposed metric, AUF_β, which evaluates the robustness to thresholding.

References

[1] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[2] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on image processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

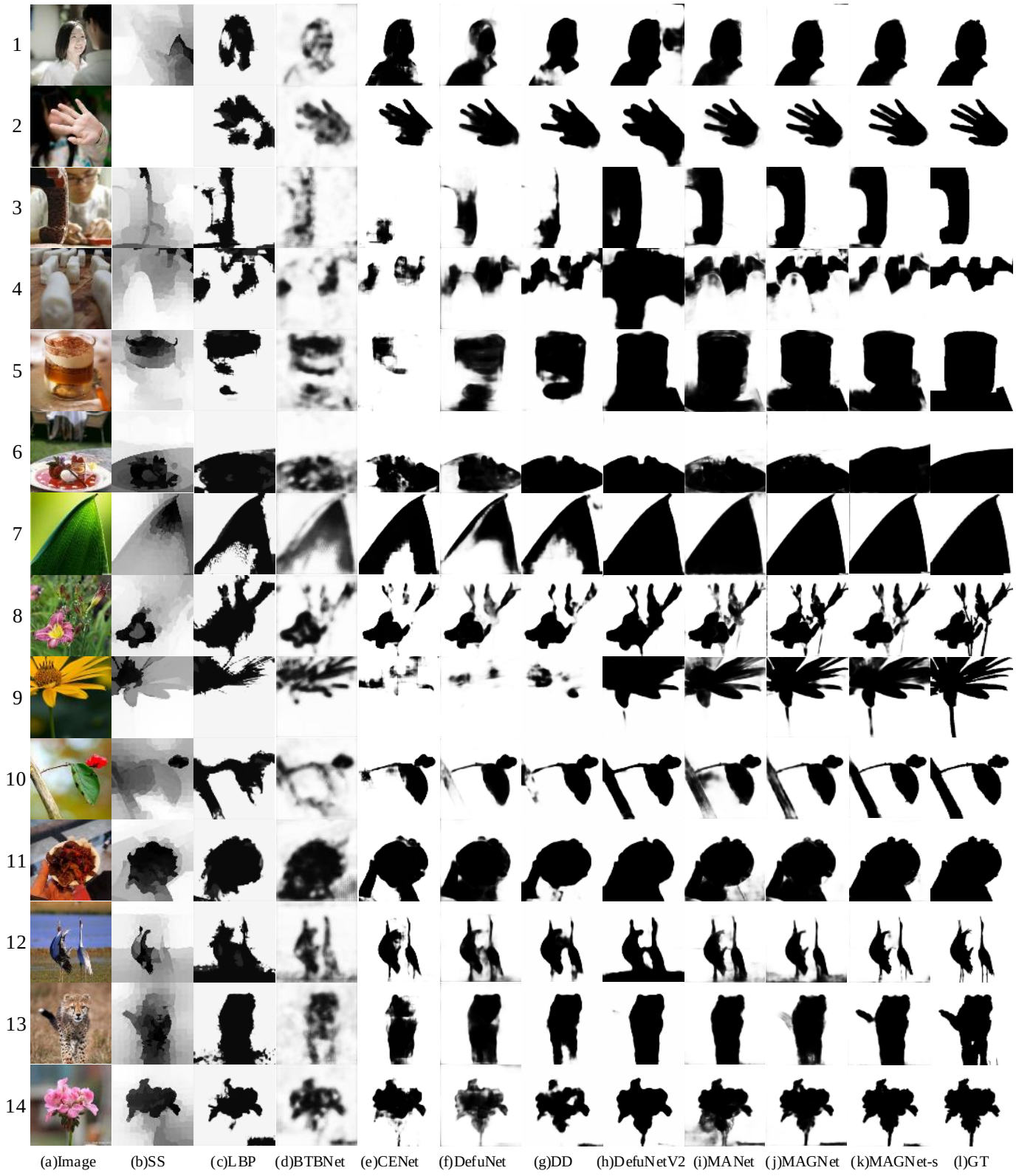


Fig. 8: Visual comparisons of the proposed method and the state-of-the-art algorithms

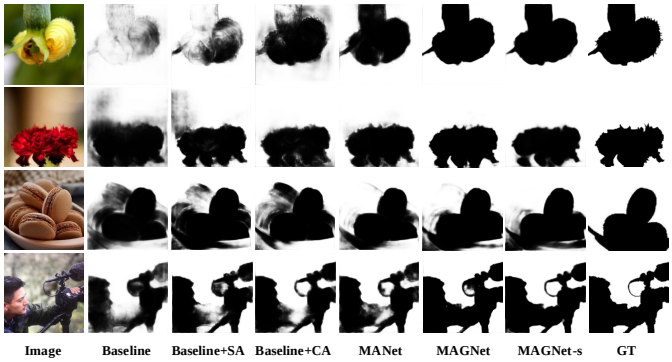


Fig. 9: Visualization of the effectiveness of different components

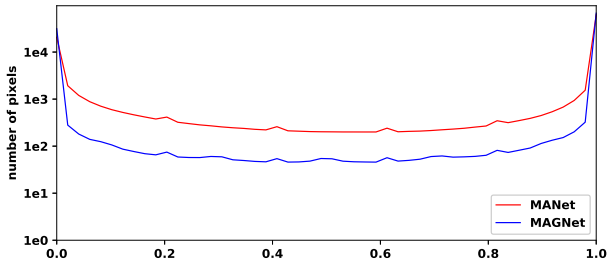


Fig. 10: Distribution of pixels predicted by MANet and MAGANet on DUT dataset.

[3] X. Wang, B. Tian, C. Liang, and D. Shi, "Blind image quality assessment for measuring image blur," in *2008 Congress on Image and Signal Processing*, vol. 1, 2008, pp. 467–470.

[4] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3085–3094.

[5] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE transactions on image processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[6] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE transactions on image processing*, vol. 25, no. 8, pp. 3919–3930, 2016.

[7] S. Bae and F. Durand, "Defocus magnification," in *Computer Graphics Forum*, vol. 26, no. 3, 2007, pp. 571–579.

[8] J. Shi, L. Xu, and J. Jia, "Discriminative blur detection features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2965–2972.

[9] A. Danielyan, V. Katkovnik, and K. Egiazarian, "Bm3d frames and variational image deblurring," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1715–1728, 2011.

[10] W. Zhang and W.-K. Cham, "Single-image refocusing and defocusing," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 873–882, 2011.

[11] X. Zhu, S. Cohen, S. Schiller, and P. Milanfar, "Estimating spatially varying defocus blur from a single image," *IEEE Transactions on image processing*, vol. 22, no. 12, pp. 4879–4891, 2013.

[12] W. Zhao, F. Zhao, D. Wang, and H. Lu, "Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3080–3088.

[13] C. Tang, X. Zhu, X. Liu, L. Wang, and A. Y. Zomaya, "Defusionnet: Defocus blur detection via recurrently fusing and refining multi-scale deep features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2700–2709.

[14] X. Yi and M. Eramian, "Lbp-based segmentation of defocus blur," *IEEE transactions on image processing*, vol. 25, no. 4, pp. 1626–1638, 2016.

[15] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[16] Y. Cao, C. Shen, and H. T. Shen, "Exploiting depth from single monocular images for object detection and semantic segmentation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 836–846, 2016.

[17] C. Tang, L. Xinwang, X. Zheng, W. Li, J. Xiong, L. Wang, A. Zomaya, and A. Longo, "Defusionnet: Defocus blur detection via recurrently fusing and refining discriminative multi-scale deep features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[18] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, pp. 2672–2680, 2014.

[21] X. Cun and C.-M. Pun, "Defocus blur detection via depth distillation," in *European Conference on Computer Vision*, 2020, pp. 747–763.

[22] Z. Jiang, X. Xu, C. Zhang, and C. Zhu, "Multianet: a multi-attention network for defocus blur detection," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1–6.

[23] Y. Pang, H. Zhu, X. Li, and X. Li, "Classifying discriminative features for blur detection," *IEEE Transactions on Cybernetics*, vol. 46, no. 10, pp. 2220–2227, 2015.

[24] B. Su, S. Lu, and C. L. Tan, "Blurred image region detection and classification," in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 1397–1400.

[25] S. A. Golestaneh and L. 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*, 2017.

[26] G. Xu, Y. Quan, and H. Ji, "Estimating defocus blur via rank of local patches," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5371–5379.

[27] K. Purohit, A. B. Shah, and A. Rajagopalan, "Learning based single image blur detection and segmentation," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 2202–2206.

[28] Y. Tang and X. Wu, "Scene text detection and segmentation based on cascaded convolution neural networks," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1509–1520, 2017.

[29] J. Park, Y. Tai, D. Cho, and I. S. Kweon, "A unified approach of multi-scale deep and hand-crafted features for defocus estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1736–1745.

[30] N. Zhang and J. Yan, "Rethinking the defocus blur detection problem and a real-time deep dbd model," in *European Conference on Computer Vision*, 2020, pp. 617–632.

[31] C. Tang, X. Liu, X. Zhu, E. Zhu, K. Sun, P. Wang, L. Wang, and A. Zomaya, "R^{2}mrf: Defocus blur detection via recurrently refining multi-scale residual features," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 063–12 070.

[32] H. Heng, H. Ye, and R. Huang, "Defocus blur detection by fusing multiscale deep features with conv-lstm," *IEEE Access*, 2020.

[33] W. Zhao, B. Zheng, Q. Lin, and H. Lu, "Enhancing diversity of defocus blur detectors via cross-ensemble network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8905–8913.

[34] C. Tang, X. Liu, S. An, and P. Wang, "Br^{2}net: Defocus blur detection via bidirectional channel attention residual refining network," *IEEE Transactions on Multimedia*, 2020.

[35] J. Lee, S. Lee, S. Cho, and S. Lee, "Deep defocus map estimation using domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 222–12 230.

[36] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[39] C. Li and M. Wand, "Combining markov random fields and convolutional neural networks for image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2479–2486.

[40] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016, pp. 694–711.

[41] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Advances in neural information processing systems*, 2017, pp. 465–476.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Machine Learning*, 2015.

[43] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 714–722.

[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[45] S. Martin Arjovsky and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the International Conference on Machine Learning*, 2017.

[46] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proceedings of the International Conference on Machine Learning*, 2018.

[47] J. Shi, L. Xu, and J. Jia, "Just noticeable defocus blur detection and estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 7132–7141.

[48] W. Zhao, X. Hou, X. Yu, Y. He, and H. Lu, "Towards weakly-supervised focus region detection via recurrent constraint network," *IEEE Transactions on Image Processing*, 2019.

[49] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[50] A. Chakrabarti, T. Zickler, and W. T. Freeman, "Analyzing spatially-varying blur," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2512–2519.

[51] C. Tang, J. Wu, Y. Hou, P. Wang, and W. Li, "A spectral and spatial approach of coarse-to-fine blurred image region detection," *IEEE Signal Processing Letters*, vol. 23, no. 11, pp. 1652–1656, 2016.