MultiANet: a Multi-Attention Network for Defocus Blur Detection

Zeyu Jiang[1], Xun Xu[2], Chao Zhang[3], Ce Zhu[1*]
[1]*School of Information and Communication Engineering, University of Electronic Science and Technology of China, China*
[2]*Institute for Infocomm Research, A-STAR, Singapore*
[3]*Sichuan Police College, China*
{jzeyu1997, alex.xun.xu, galoiszhang}@gmail.com, eczhu@uestc.edu.cn

***Abstract*—Defocus blur detection is a challenging task because of obscure homogenous regions and interferences of background clutter. Most existing deep learning-based methods mainly focus on building wider or deeper network to capture multi-level features, neglecting to extract the feature relationships of intermediate layers, thus hindering the discriminative ability of network. Moreover, fusing features at different levels have been demonstrated to be effective. However, direct integrating without distinction is not optimal because low-level features focus on fine details only and could be distracted by background clutters. To address these issues, we propose the Multi-Attention Network for stronger discriminative learning and spatial guided low-level feature learning. Specifically, a channel-wise attention module is applied to both high-level and low-level feature maps to capture channel-wise global dependencies. In addition, a spatial attention module is employed to low-level features maps to emphasize effective detailed information. Experimental results show the performance of our network is superior to the state-of-the-art algorithms.**

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 widespread applications including but not limited to foreground segmentation, blur magnification, all-in-focus image generation and depth estimation.

Traditional defocus blur detection methods usually use the low-level features such as gradient and frequency features [1], [2] to extract the boundaries. Despite significant progress, these hand-crafted based methods are usually effective only in limited scenes where the boundary is clear enough to separate in-focus and out-of-focus (blurred) regions. Therefore, they may fail when trying to separate the smooth blurred region which do not contain obvious boundary from the smooth in-focus region which is also referred to as homogeneous area. To address these issues, deep Convolution Neural Networks (DCNNs) have been applied to defocus blur detection tasks recently. Zhao et al. [3] proposed the BTBNet which integrates the semantic cues and structural information by designing a multi-stream fully convolution network and resolves the misclassification of in-focus homogeneous areas. Tang et al. [4] proposed the DefusionNet which uses recurrent fusion and refine module to integrate multi-level information. Although these methods have made significant progress, their networks do not explicitly consider the interdependencies between features in either spatial or feature channel dimensions. In addition, they fuse all the detailed features without considering their different contributions to defocus blur detection. Hence, their results are sometimes interfered by the cluttered background, as shown in Fig. 1 and some low-contrast focal regions are misclassified.

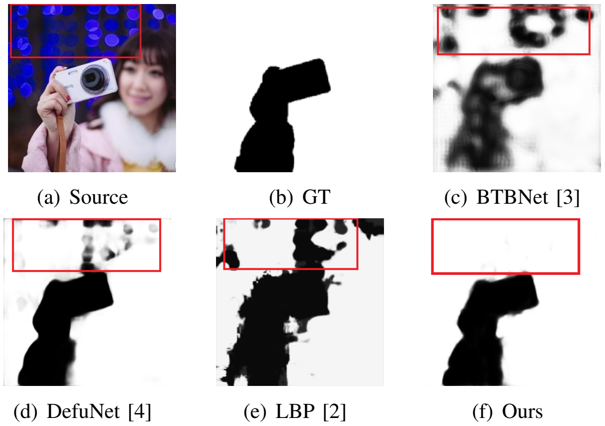
(a) Source (b) GT (c) BTBNet [3]
(d) DefuNet [4] (e) LBP [2] (f) Ours

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 complementary for respective tasks, where the former captures the global context information and the later captures the spatial structural details [5], [6]. Both the high-level and low-level information are further integrated for better feature representation by attention mechanism. Therefore, we believe it is beneficial to exploit the separation of high-level and low-level information and introduce attention mechanisms to fuse features. Specifically, we propose a novel defocus blur detection method named Multi-Attention Network, which is illustrated in Fig. 2.

In order to improve the discriminative learning ability of network, we first introduce a channel-wise attention module to explicitly model interdependencies of features by calculating 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 with [4] 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 awkward with homogeneous areas, we further propose to use the high-level information to guide the learning of low-level features by providing spatial cues. For this purpose, we introduce a spatial attention module to guide low-level features with a spatial attention map computed 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.

Our main contributions can be summarized as follows:

- A novel Multi-Attention Network (MultiANet) 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.
- In order to enhance the discriminative ability of network, a channel-wise attention is performed to explicitly capture interdependencies between layers. Moreover, a spatial attention is employed to extract desired details and suppress the interference from background clutter. To the best of our knowledge, this work is the first attempt to exploit an end-to-end deep network combining channel-wise attention and spatial attention for defocus blur detection.
- We report the state-of-the-art on two benchmark datasets. Our method consistently outperforms other state-of-the-art methods through extensive experiments.

II. Related Work

Defocus Blur Detection (DBD). The traditional methods are based on hand-crafted features and mostly focus on the differences of gradients and frequency information between in focus and out-of-focus regions. Golestaneh and Karam [7] 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 [2] 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. [8] introduced a ranking-based metric to detect defocus blur regions. 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 are unsatisfying when dealing with complex scenes.

Deep CNNs could effectively extract semantic features and by combining the multi-level features, the performance of DBD methods has been improved significantly. Park et al. [9] introduced a unified approach to combine handcrafted and deep features to detect out-of-focus regions. Zhao et al. [3] proposed a multi-stream bottom-top-bottom fully convolutional network to extract more features by constructing wider and deeper network. However, their large number of parameters lead to high storage and computation consumption. Tang et al. [4] proposed a defocus blur detection method based on recurrently fusing and refining the feature maps. Zhao et al. [10] introduced a cross-ensemble network to enhance diversity of defocus blur detectors. Tang et al. [11] proposed a bidirectional residual refining network for blur detection.

Despite the improvement these methods have made, they neglect to extract the correlations of feature maps and integrate all detailed features without distinction. Thus their results sometimes are interfered by the background clutter (as shown in Fig. 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. Wang et al. [12] proposed the non-local network mainly exploring effectiveness of non-local operation in spacetime dimension for videos and images. Zhao et al. [5] proposed a pyramid feature attention work for saliency detection. However, their attention module neglects to take into consideration the relationship of feature maps, which is crucial for enhancing the discriminative ability of network. Fu et al. [6] designed two parallel self-attention modules to capture long-range dependencies for semantic segmentation task. Different from previous works, we extend the attention mechanism to the task of defocus blur detection and design a Multi-Attention Network which could not only learn better feature representation, but also adaptively filter out noise from background. Comprehensive experiment results demonstrate the effectiveness of our method.

III. Method

A. Overview of the MultiANet

Most of 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 misclassifying homogeneous areas. To resolve these issues, we develop an efficient defocus blur detection network taking into consideration the correlation between feature layers and use the spatial attention of high-level features to guide low-level feature learning. As illustrated in Fig. 2, the pre-trained model VGG-16 is employed as the backbone, and we divide the layers into two groups. Specifically, conv3_3, conv4_3 and conv5_3 are deeper layers to exploit high-level features. For the deep layers, we up-sample the conv4_3 and con5_3 to the size of conv3_3, then combine them by a cross channel concatenation as the basic high-level features. Meanwhile, conv1_2 and conv2_2 are shallow layers 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

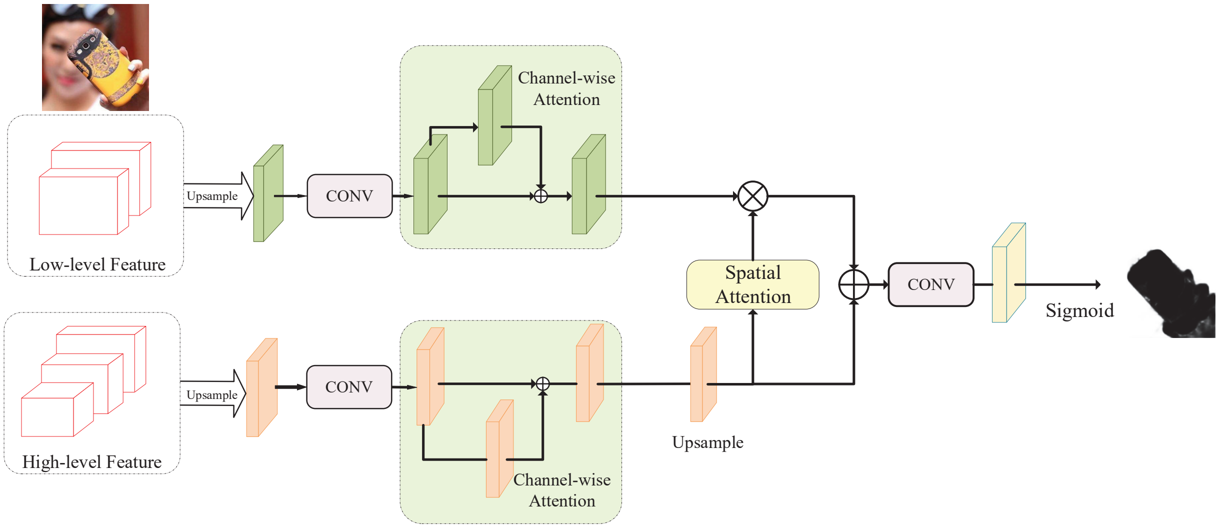

Fig. 2: Framework of the proposed Multi-Attention Network (MultiANet).

different feature maps. After that, 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 are focused on detailed boundaries and are prone to background clutters [13], [14]. Finally, we fuse the output from high-level and low-level features to obtain better pixel-level predictions.

B. Channel attention module

Existing deep learning based methods focus mainly on designing a deeper or wider network to learn more discriminative high-level features, while rarely exploiting the inherent feature correlations in intermediate layers, thus hindering the representational ability of CNNs. To address the issue, 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 Fig. 3, given a feature $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$,as [6], [12], we firstly reshape it to $\mathbb{R}^{C\times N}$, then perform a matrix multiplication between $\mathbf{X}$ and its transpose. After that, the attention map $\mathbf{A}$ is obtained by performing a softmax layer,

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

where a_{ji} 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 $\mathbf{X}$ and $\mathbf{A}$ 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 $\mathbf{Y}$.

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

The scale factor α is initialized to zero which means the module have no influence on the input feature maps at first and gradually learns a proper weight during the training process. It can be inferred from (2) that the resulting features map $\mathbf{Y}$ is a weighted sum of all channels and the original map. Therefore, it models the interdependencies cross 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.

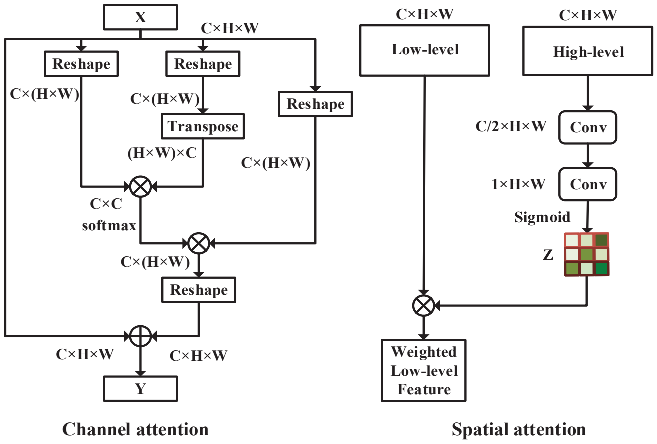

Fig. 3: The details of channel attention module (left) and spatial attention module (right).

C. Spatial attention module

The low-level cues are essential to defocus blur detection for helping 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 or even misclassification. For instance, some out-of-focus regions with strong detailed information may be mistakenly regarded as in-focus regions as Fig. 1. To address this issue, we propose a spatial attention module to adaptively emphasize on desired low-level features. As illustrated in Fig. 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, $\mathbf{H} \in \mathbb{R}^{C\times H\times W}$ stands for high-level features and $\mathbf{L} \in \mathbb{R}^{C\times H\times W}$ stands for low-level features. In order to increase receptive field without additional computation cost, two consecutive atrous convolutions are applied to extract spatial information (see Fig. 3). 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 $\tilde{\mathbf{L}}$ is acquired by weighting low-level feature $\mathbf{L}$ with spatial attention weight map $\mathbf{Z}$ as,

$$\tilde{\mathbf{L}} = \mathbf{L} \cdot \mathbf{Z} = \mathbf{L} \cdot f(\mathbf{H}, \mathbf{W}) \tag{3}$$

where W refers to parameters in spacial attention block. It can be inferred from (3) that low-level features are explicitly rescaled by the spatial attention module. Large weight would be assigned to the low-level features which play important role in refining the boundaries. Hence, according to high-level semantic cues, the spatial attention module could focus more on effective details and filter out some background clutter.

IV. Experiment

A. Datasets and Implementation Details

Datasets. In our experiments, we use two publicly available datasets with pixel-level annotations. **Shi's Dataset** 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 last 100 images for testing as [4]. **DUT** is a new dataset for defocus blur detection proposed by [3] which consists of 600 training images and 500 test images with pixel-level annotations. It is a more challenging datasets because images have multiscale focused area, low contrast focal regions and strong background clutter.

Implementation Details. The overall architecture is illustrated in Fig. 2. Our network uses the VGG-16 [15] as backbone. The cross-entropy loss is applied to the output of this network. Two public datasets have limited number of training sets which is insufficient to train a deep neural network. In order to improve generalization and reduce overfitting, we utilize data augmentation to the origin images by randomly flipping, cropping, resizing and rotating. Our model is trained on the Nvidia GTX Titan Xp. The whole network is optimized by Adam algorithm and the learning rate is initialized to 0.0004. The training batch size is set to 6 and the whole training process on Shi's datasets takes roughly 3 hours.

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

We compare our method with other 8 state-of-the-art approaches, including ASVB [16], DBDF [1], SS [17], LBP [2], HiFST [7], BTBNet [3], DeFusionNet [4] and BR2Net [11]. Same as other state-of-the-art methods [4], four widely used metrics are used to quantitatively evaluate the performance of the proposed network: F-measure curves, mean absolute error (MAE), F-measure score (F_β) and precision-recall curves. As to the last three methods, we directly copy the results from the authors' project page.

Quantitative Comparison. Tab. I and Fig. 4 provide the quantitative evaluation results of the proposed method and eight state-of-the-art defocus blur detection approaches in terms of PR curve, F-measure curve, F_β and MAE criteria. It is observed that our method consistently outperforms our counterparts in terms of F_β and MAE criteria, which demonstrates the effectiveness of our method. In particular, our method gets larger improvement compared with the best existing approach on DUT dataset. DUT dataset is a difficult and challenging defocus blur detection dataset, which contains many complex natural scenes images and strong background clutter. The proposed method can effectively enhance the discriminative ability of network and obtain desired detail features, therefore yields better detection results.

Running Efficiency Comparison. Apart from the fine results, our method is also efficient. We use one GPU(Nvidia GTX Titan Xp) in both training and testing process. The average running time for an image of different methods are shown in Tab. II. Our method is faster than other methods for detecting the out-of-focus regions.

Visual Comparison. Fig. 5 provides a visual comparison of our method and other state-of-the-art approaches. We observe that our method gest the best detection results. Specifically, the background clutter are effectively suppressed. By enhancing the discriminative ability of network, the low-contrast focal regions can be accurately classified. With the help of the combination of attention modules, desired structural information has been extracted and the network yields much more clear boundaries.

C. Ablation Study

Effectiveness of Attention Modules. To analyze the impact of different modules in our network, we carry out ablation studies on the DUT dataset which has more complex scenes and more images. For fair comparison, we perform the same data augmentation operations on all networks. In order to evaluate our Multi-Attention Network (MultiANet), we use the high-level features from VGG as the baseline. We denote HL, LL, HCA, LCA and SA for High-level features, low-level features, high-level channel-wise attention module, Low-level channel-wise module and spatial attention module respectively.

TABLE I: Quantative comparison of F-measure and MAE scores. The best two results are shown in red and blue colors, respectively.

Datasets	Metric	ASVB [16]	DBDF [1]	SS [17]	LBP [2]	HiFST [7]	BTBNet [3]	DeFuNet [4]	BR2Net [11]	Ours
Shi's	F_β	0.731	0.841	0.787	0.866	0.865	0.892	0.917	0.918	0.951
	MAE	0.636	0.323	0.298	0.186	0.232	0.105	0.116	0.132	0.096
DUT	F_β	0.747	0.802	0.784	0.874	0.866	0.887	0.922	0.943	0.950
	MAE	0.651	0.369	0.296	0.173	0.302	0.190	0.115	0.104	0.078

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

	Methods	ASVB	DBDF	SS	LBP	HiFST	BTBNet	DeFuNet	BR2Net	Ours
Datasets	Shi's	2.04	214.83	2.76	57.34	2576.24	25	0.094	0.087	**0.062**
	DUT	1.59	110.37	1.20	30.38	1169.57	25	0.056	0.053	**0.050**

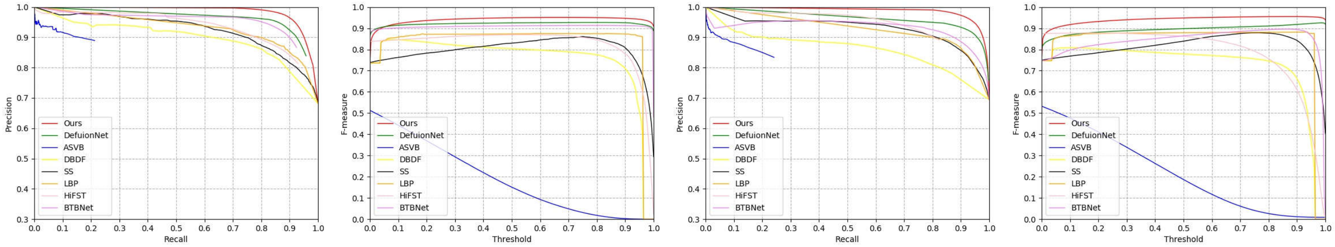

Fig. 4: Quantitative comparisons of the proposed approach and 8 state-of-the-art approaches on two datasets. The first two are the PR curves and F-measure curves of different methods on Shi's dataset respectively. The last two are the PR curves and F-measure curves of different methods on DUT dataset respectively.

Results are reported in Tab. III. Obviously, the proposed network with all the components yields the best results on both MAE and F_β and each of the component is contributing positively to the overall framework.

Visualize Channel Attention Effectiveness. To further demonstrate the impact of the channel-wise attention module we compare the results of MultiANet without high-level channel attention module and the proposed MultiANet with results shown in Fig. 6. Apparently, some low contrast in-focus regions are misclassified (first row) and the cluttered background interferes the detection result (second row). This is because the network without channel attention is unable to extract the inter-dependencies of features thus hindering the discriminative ability.

Visualize Spatial Attention Effectiveness. Similarly, we visualize the impact of spatial attention module by comparing the propose network without spatial attention with the fully MultiANet. As shown in Fig. 7, the network without spatial attention is unable to adaptively select correct spatial extent, hence, its detection results are influenced by the homogeneous areas, e.g. the smooth out-of-focus areas on macarons.

TABLE III: Ablation analysis of the different components combinations.

Methods	MAE	F_β
VGG HL(Baseline)	0.132	0.924
VGG HL+LL	0.101	0.937
VGG+HL+LL+HCA	0.092	0.938
VGG+HL+LL+HCA+SA	0.088	0.944
VGG+HL+LL+HCA+LCA	0.089	0.938
VGG+HL+LL+LCA+SA	0.087	0.940
VGG+HL+LL+HCA+LCA+SA	**0.078**	**0.950**

V. Conclusion

We propose a novel method named Multi-Attention Network (MultiANet) 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. Finally, MultiANet obtains the final defocus blur map by fusing low-level and high-level features. Extensive experimental results demonstrate our method outperforms other state-of-the-art methods in terms of both accuracy and efficiency.

References

[1] J. Shi, L. Xu, and J. Jia, "Discriminative blur detection features," in *CVPR*, 2014.

[2] X. Yi and M. Eramian, "Lbp-based segmentation of defocus blur," *IEEE Trans. Image Processing*, 2016.

[3] W. Zhao, F. Zhao, D. Wang, and H. Lu, "Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network," in *CVPR*, 2018.

[4] 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 *CVPR*, 2019.

[5] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *CVPR*, 2019.

[6] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *CVPR*, 2019.

[7] S. A. Golestaneh and L. J. Karam, "Spatially-varying blur detection based on multiscale fused and sorted transform coefficients of gradient magnitudes," in *CVPR*, 2017.

[8] G. Xu, Y. Quan, and H. Ji, "Estimating defocus blur via rank of local patches," in *ICCV*, 2017.

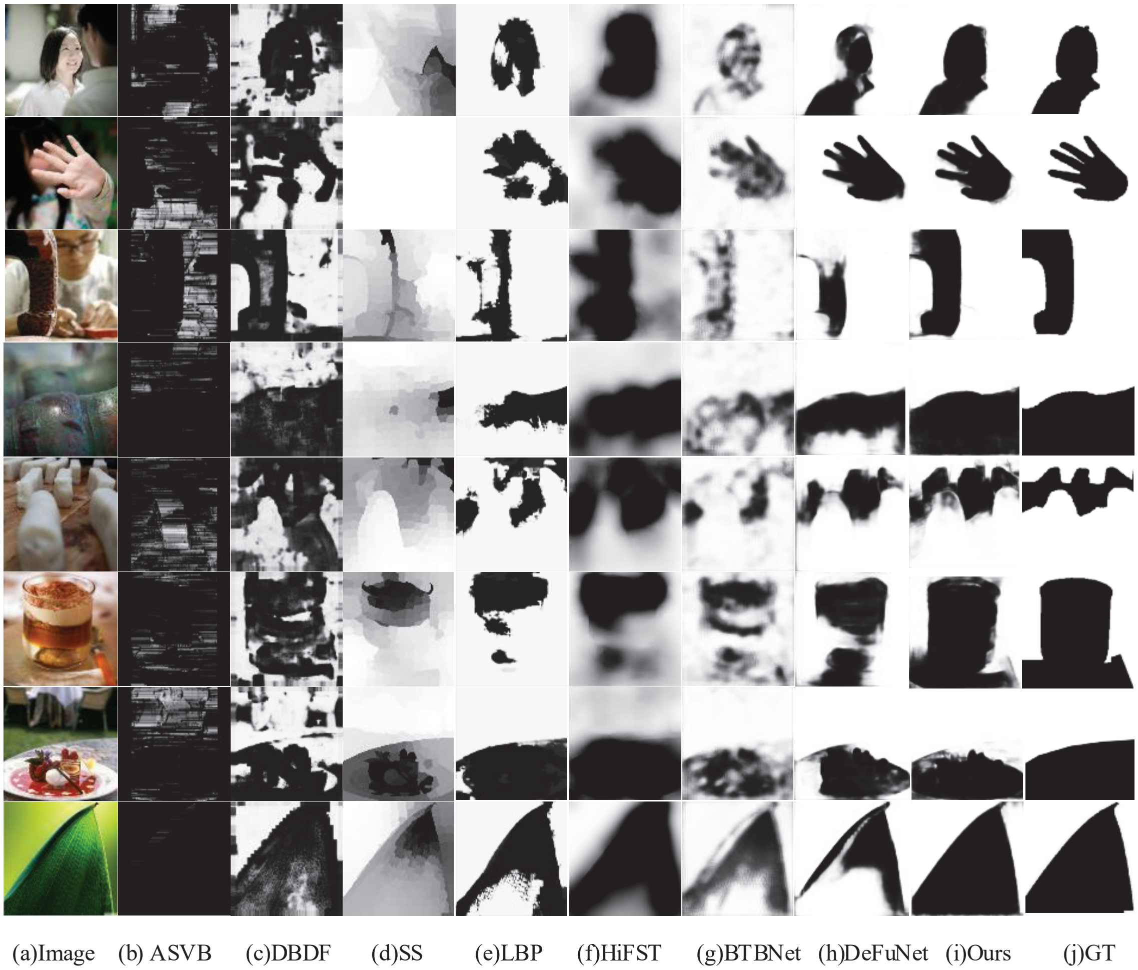

Fig. 5: Visual comparisons of the proposed method and the state-of-the-art algorithms.

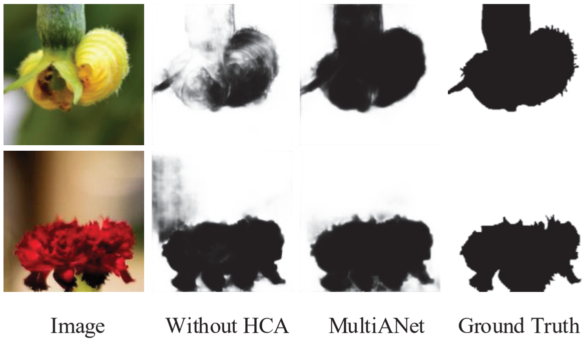

Fig. 6: Visualization of the effectiveness of channel-wise attention.

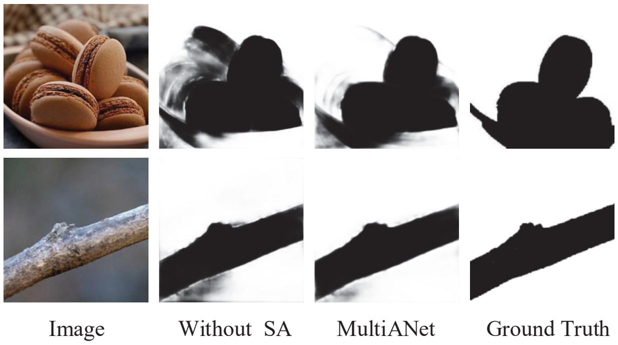

Fig. 7: Visualization of the effectiveness of spatial attention.

[9] 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 *CVPR*, 2017.

[10] W. Zhao, B. Zheng, Q. Lin, and H. Lu, "Enhancing diversity of defocus blur detectors via cross-ensemble network," in *CVPR*, 2019.

[11] 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.

[12] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018.

[13] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *CVPR*, 2018.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[16] A. Chakrabarti, T. E. Zickler, and W. T. Freeman, "Analyzing spatially-varying blur," in *CVPR*, 2010.

[17] 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 Process. Lett.*, 2016.