This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2021.3128375, IEEE Signal Processing Letters

Heterogeneous Attention Nested U-Shaped Network for Blur Detection

Wenliang Guo, Xiao Xiao, *Member, IEEE*, Yilong Hui, *Member, IEEE*, Wenming Yang, *Senior Member, IEEE*, and Amir Sadovnik, *Member, IEEE*

***Abstract*—With the popularity of image sensors in various mobile devices, image blurring caused by hand shaking or out of focus becomes ubiquitous, which deteriorates image quality and poses challenges for vision tasks, including object detection, image classification and image segmentation. Designing an efficient blur detection algorithm which can automatically detect and locate blurred regions becomes necessary. In this letter, we design an end-to-end convolution neural network called heterogeneous attention nested U-shaped network (HANUN) for blur detection. We introduce pyramid pooling into encoders to enhance the feature extraction at different scales and reduce the gradual information loss. Inspired by the nested network design, small U-shaped networks are embedded into our decoders to increase the network depth and promote feature fusion with different receptive field scales. In addition, we incorporate a channel attention mechanism in the proposed network to highlight the informative features for detecting the blurry regions. Experimental results show that HANUN outperforms other state-of-the-art algorithms for blur detection tasks on public datasets and real-world images.**

***Index Terms*—Blur detection, convolution neural network, heterogeneous design, nested structure, semantic information**

I. Introduction

NOWADAYS, with the widespread using of mobile devices, image blurring becomes ubiquitous. Image blurring can degrade the quality of images, which poses challenges for the subsequent image processing, especially in computer vision applications. Before the emergence of deep learning, blur detection was mainly achieved by using traditional machine learning algorithms [1]–[8]. Since the features contained in the blurred regions vs. the sharp regions are different, most of the traditional methods begin by comparing or distinguishing the features of different image regions in special domains, such as the frequency domain. For example, Tang *et al.* [5] find that the spectrum amplitude can be affected by defocusing blur at edges, and they conduct blur detection through establishing the relationship between blurred regions and the image spectrum. Shi *et al.* [6] construct a blur detection algorithm which combines several image characteristics, including image gradients, frequency-domain spectrum and local filters.

Although traditional methods provide preliminary solutions to blur detection tasks, it is still a challenge to separate sharp regions from blurry ones due to the fixed hand-crafted features, which may not be ideal and the lack of high-level semantic feature extraction [9]. In recent years, we have witnessed the great success of convolutional neural networks (CNNs) in many fields [10]–[18], including blur detection [19], [20]. Furthermore, inspired by image segmentation, some segmentation networks [21]–[23] and design ideas, such as auto encoding [24]–[26] as well as the extraction of high-level semantic information [9], are applied to blur detection [27], [28].

As the U-shaped architecture is widely utilized in segmentation networks [21], [23], it also demonstrates a great advantage for blur detection [28], [29]. In particular, U^2-Net firstly utilize the nested structure, which inspires our work. However, since most existing U-shaped networks utilize homogeneous architecture, i.e., the symmetric encoder-decoder structures, their encoding and decoding ability is greatly constrained. In addition, the features generated by the encoder and decoder are directly combined by a skip connection, which is not reasonable enough because the differences in the information contained in the encoded and decoded features, and the proportion of the contribution of the different features to the prediction results are not taken into account.

In this letter, we propose an end-to-end CNN called heterogeneous attention nested U-shaped network (HANUN) for blur detection. We nest pyramid poolings into encoders to effectively extract and fuse features at different scales. Instead of using a series of convolutions, we utilize pyramid poolings because it can reduce the general information loss in cascade layer design. We nest U-shaped networks into decoders to increase the depth, which improves the network robustness and generalization capability without significantly increasing the number of parameters [23]. We also introduce an attention mechanism to augment the informative features. Our nested and heterogeneous designs enable HANUN to show high performance on public datasets and real-world images.

The rest of this letter is organized as follows. Section II introduces our network design. Section III tests the performance of the proposed network on public datasets and our real-world dataset. Section IV performs ablation studies to explore the effect of each design to the network's performance. Section V summarizes the letter with the conclusion.

W. Guo, X. Xiao and Y. Hui are with the school of Telecommunications Engineering, Xidian University, Xi'an 710071, China e-mail: (wlguo@stu.xidian.edu.cn; xiaoxiao@xidian.edu.cn; ylhui@xidian.edu.cn).

W. Yang is with the Shenzhen Key Laboratory of Information Science and Technology, Shenzhen International Graduate School, Tsinghua University, Beijing 100084, China (e-mail: yangelwm@163.com).

Amir Sadovnik is with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996, USA email: (asadovnik@utk.edu).

This work was supported in part by the NSFC under Grant 61901341 and 61403291, in part by the China Postdoctoral Science Foundation under Grant 2021TQ0260, and in part by the National Natural Science Foundation of Shaanxi Province under Grant 2020JQ-301.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2021.3128375, IEEE Signal Processing Letters

II. Proposed Network

In this section, we propose our blur detection network called heterogeneous attention nested U-shaped network (HANUN) shown in Figure 1. We nest a pyramid pooling module [22] in the encoder which we call pyramid pooling encoder (PP encoder), and we nest an attention block as well as a nested U-shaped network in the decoder which we call attention nested U-Net decoder (ANU decoder). The heterogeneous design (i.e. asymmetric design of the network encoding and decoding sides) alleviates the difficulty of achieving optimal performance of both encoders and decoders in homogeneous design. It enables semantic and contextual information at more scales to be extracted and fused, which further improves the performance of the proposed network HANUN. In this work, HANUN contains five encoders and six decoders in total. Based on the positions in the network, we categorize the encoders and decoders into shallow and deep ones, and we employ different designs to each category.

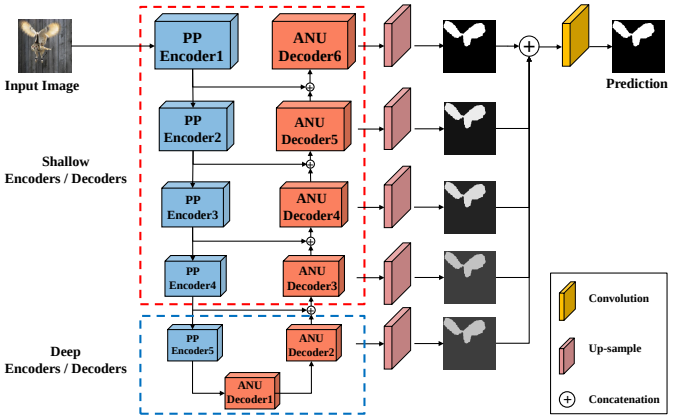


Fig. 1: The architecture of the heterogeneous attention nested U-shaped network (HANUN).

A. Pyramid Pooling Encoder

The PP encoder consists of pyramid pooling and some subsequent operations, while the former is the core of our encoder. A pyramid pooling layer contains multiple pooling layers with different pyramid pooling scales. As shown in Figure 2, pooling layers with different scales are drawn in different colors. Each pooling layer will generate a sub-feature, hence the receptive field scales of the sub-features are all different. Compared with a series of convolutions, pyramid pooling uses a parallel layer structure instead of cascade, and all pooling operations are performed on the input feature directly. Therefore, there is less original information loss in the high-level features with large receptive fields. Moreover, an encoder using pyramid poolings has fewer parameters than using convolutions, which is verified in Section IV.

Afterward, each sub-feature goes through a bottleneck layer to change its dimension. Then the features are up-sampled to their original size (their size before pooling), and concatenated with the input feature. Through these connections on the channel dimension and subsequent convolutions, the semantic

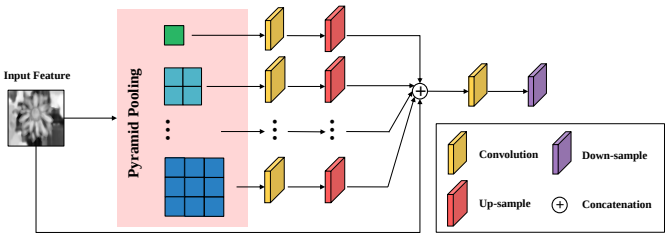


Fig. 2: The structure of pyramid pooling encoder (PP encoder).

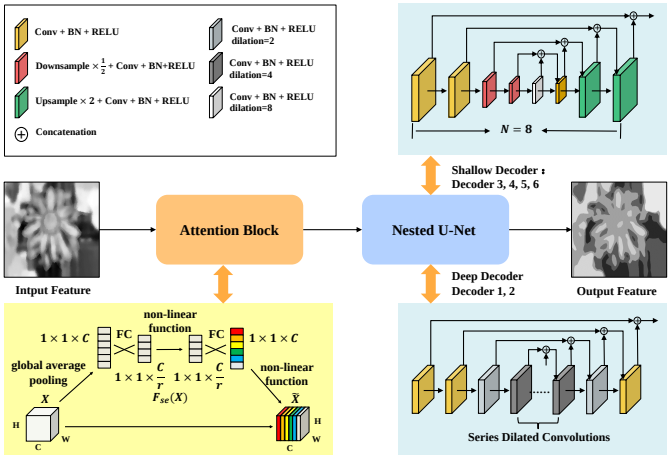


Fig. 3: The structure of the attention nested U-Net decoder (ANU decoder).

and contextual information contained in the different sub-features are fused.

B. Attention Nested U-Net Decoder

In order to solve the problem of the difference in the feature information and the contribution to the result in the concatenation process, we introduce attention mechanism by exploiting an SE-block [30] (shown in Figure 3) after feature combination into the ANU decoder. The SE-block can model the spatial dependence of features and realize channel attention through recalibrating the channel-wise feature responses. Accordingly, by using the attention block in HANUN, the importance of different features are distinguished through training, and the informative features for prediction can be augmented.

From the perspective of feature sizes, since they are quite different in each decoding stage, we present different feature processing designs to shallow and deep decoders. In deep decoders shown in the lower right corner of Figure 3, the middle parts are a series of dilated convolutions with different dilation rates. In our design, the feature sizes after each dilated convolution layer are all the same. Given that the feature size is relatively small in the deep decoders, this design can effectively reduce the loss of useful information due to further shrinking the feature size.

Shallow decoders are shown in the upper right corner of Figure 3. Since the sizes of features are relatively large in them compared with the aforementioned deep decoders, we nest a normal U-shaped convolutional network (U-Net) with down-samplings to expand the receptive fields and extract

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2021.3128375, IEEE Signal Processing Letters

TABLE I: Quantitative comparisons on the CUHK dataset [6].

Metric	DBDF [6]	JNB [7]	BTBNet [27]	LBP [4]	CENet [31]	U^2-Net [23]	**HANUN**
$F_{\sqrt{0.3}}$	0.6438	0.7669	0.8491 (0.867)	0.8688 (0.864)	0.8757 (0.906)	0.9247	**0.9701**
MEA	0.3402	0.3537	0.0820 (0.107)	0.2618	0.0593 (0.059)	0.0980	**0.0364**

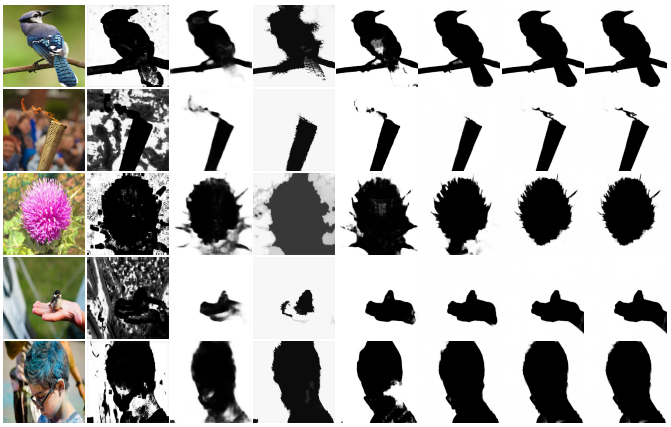


Fig. 4: Comparison on testing results with other state-of-the-art networks.

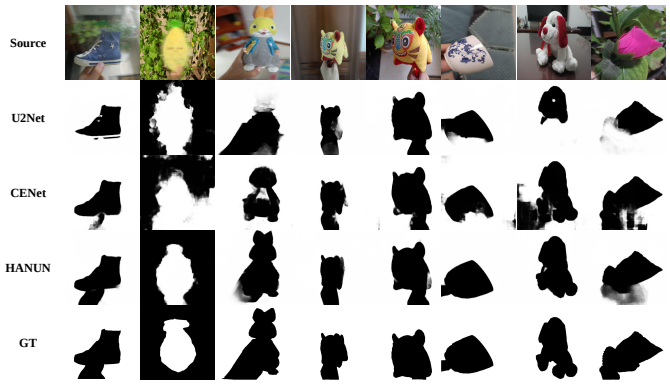


Fig. 5: Result examples on the real-world image dataset using the different networks.

high-level features. However, in each decoding stage, while the features reach the middle layers of the U-Net, their sizes are continuously reduced due to down-samplings. Hence, in order to prevent the intermediate features from being too small, which may also lead to the loss of potential informative features, we restrict the number of layers (represented as N in the figure) in each shallow decoder by hierarchically reducing it. In this work, the values of N are set to 8, 10, 12 and 14 from decoder 3 to decoder 6, respectively. This design effectively avoids the occurrence of the too-small intermediate feature, enabling HANUN to adapt to small-scale input images.

III. Experiments

In this section and Section IV, two public datasets: CUHK [6] and DUT [27] are used to evaluate the performance of HANUN. CUHK dataset contains 1000 blurred images in total, including 296 motion-blurred images and 704 defocus-blurred images. DUT dataset contains 500 defocus-blurred images. On public datasets, we compared HANUN with other state-of-the-art blur detection algorithms, including BTBNet [27], CENet [31], DBDF [6], JNB [7], LBP [4] and U^2-Net [23]. All the comparisons in our letter are based on released code or the results shown in the original papers.

TABLE II: Quantitative comparisons on the real-world image dataset.

Metric	CENet	U^2-Net	**HANUN**
$F_{\sqrt{0.3}}$	0.6804	0.7236	**0.8288**
MEA	0.1340	0.1680	**0.0811**

The F-measure and mean absolute error (MAE) are utilized for quantitative evaluations. The F-measure is the weighted harmonic mean of precision and recall. It is defined as:

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

where β is usually set to $\sqrt{0.3}$ in blur detection tasks. A higher F-measure reflects a better performance of a network. MAE is an indicator used to describe the error between predicted values and true values. It is defined as:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - G(x,y)|, \tag{2}$$

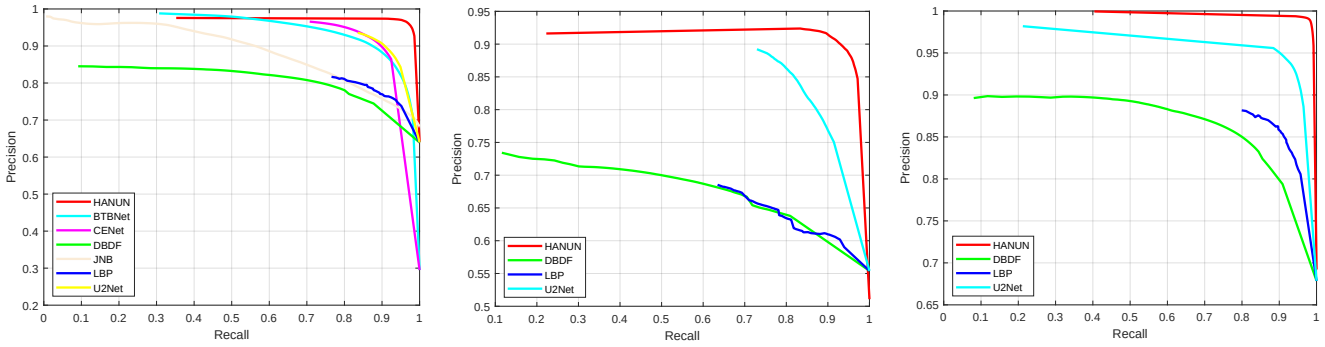
where W and H are the height and width of an image, respectively. S represents the segmentation result and G represents the ground truth (GT). A smaller MAE reflects a better performance of the network.

Figure 4 presents some detection examples of HANUN and the other networks. The results show that, compared with other networks, the targets segmented by HANUN are more complete, and the edges are clearer and sharper. The results of HANUN are the closest to the ground truth (GT).

PR curves are displayed in Figure 6. Figure 6(a) shows the testing result on the CUHK dataset. Figure 6(b) and Figure 6(c) show the testing results on motion-blurred images and defocusing-blurred images, respectively. Since BTBNet, CENet and JNB do not provide their code or their testing results divided to the two blur types, they are not shown in Figure 6(b) and Figure 6(c). From Figure 6 we can see that HANUN achieves superior results as compared with the other state-of-the-art networks, especially on motion-blurred images.

The quantitative evaluation results are shown in Table I. The data in parentheses are the results provided in the cited papers while the others are the best we are able to achieve in our experiments. We can see that our network achieves significant improvement on the blur detection performance.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2021.3128375, IEEE Signal Processing Letters



(a) All images. (b) Motion-blurred images. (c) Defocusing-blurred images.

Fig. 6: Precision-recall curves of the different networks on the CUHK dataset.

TABLE III: Results of ablation study on different components and network depth.

Network	PP Module	Nested U	SE-block	CUHK [6]		DUT [27]		Parameters
				$F_{\sqrt{0.3}}$	MAE	$F_{\sqrt{0.3}}$	MAE	
HANUN	✓	✓	✓	0.9701	0.0364	0.9205	0.1072	29.75 M
HANUN-5				0.9347	0.0498	0.9115	0.1191	8.17 M
HANUN-7				0.9723	0.0358	0.9211	0.1086	52.04 M
HANUN-allPP	✓	×	✓	0.9125	0.1895	0.9327	0.1738	25.58 M
HANUN-noSE	✓	✓	×	0.8726	0.1315	0.9096	0.1198	29.44 M
U^2-Net	×	✓	×	0.9247	0.0980	0.9328	0.0941	44.01 M

In order to show that our model can generalize to images taken by mobile-phones as opposed to only images from public datasets, we make a small real-world dataset, including 60 motion-blurred images and 40 defocus-blurred images. All images are shot by an iPhone12 Pro Max with a Sony IMX603 camera sensor. We choose U^2-Net [23] trained on CUHK, and CENet [31] trained on DUT (officially provided) for performance comparison, because they show top performance in the aforementioned evaluations.

We select several images in various scenes and show them in Figure 5. It is clear that the results of HANUN have the sharpest edges and the most complete objects, and are the closest to the ground truth (GT). The quantitative comparison results are illustrated in Table II where we see that HANUN has the best performance. The experiment on the real-world image dataset demonstrates that HANUN has high robustness and can generalize well.

IV. Ablation Studies

We build a new network called HANUN-allPP and use U^2-Net [23] for comparison to verify the effectiveness of the heterogeneous structure. HANUN-allPP replaces the nested U-Net in each decoder with pyramid pooling in the same manner as it is done in the same-depth encoder. In the U^2-Net, all encoders and decoders are based on convolutional layers. Therefore, both U^2-Net and HANUN-allPP are homogeneous networks. Experimental results in Table III shows that HANUN has the best performance on the CUHK and DUT datasets. Especially on motion-blurred images, HANUN shows the great advantage. These results demonstrate the superiority of our heterogeneous structure.

In order to verify the importance of the SE-block used to provide channel attention, we build HANUN-noSE without SE-blocks in the decoders. Table III shows that removing the SE-block causes significant performance deterioration for the network. However, compared with on the CUHK dataset, the performance reduction on the DUT dataset is not as obvious. This hints that the SE-block is one of the reasons that HANUN significantly outperforms other methods when dealing with motion-blurred images.

For exploring the optimal depth, we constructed HANUN-5 and HANUN-7 with the depth 5 and 7, respectively. Table III shows that increasing the network depth can improve the network performance. However, compared with the slight performance improvement, the number of parameters dramatic increases. As a result, taking into account both the performance limitations of the operating device itself and blur detection performance, HANUN with depth 6 (i.e., the network proposed in this letter) appears the most cost-effective.

V. Conclusion

This letter presents an end-to-end convolution neural network for blur detection. The novelty of our network is the heterogeneous and nested design. Pyramid pooling is incorporated in the encoder to maximize semantic and contextual information extraction. Our decoder decodes the features from different scales using a nested U-shaped network, which significantly enhances the robustness of the proposed model. A channel attention mechanism is also introduced into our network to augment the informative intermediate features. Experimental results show that our network outperforms other state-of-the-art networks in the blur detection tasks on public datasets and our real-world dataset.

REFERENCES

[1] S. Alireza 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, pp. 5800–5809.

[2] A. Karaali and C. R. Jung, "Edge-based defocus blur estimation with adaptive scale selection," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1126–1137, 2017.

[3] 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.

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

[5] C. Tang, C. Hou, and Z. Song, "Defocus map estimation from a single image via spectrum contrast," *Optics letters*, vol. 38, no. 10, pp. 1706–1708, 2013.

[6] 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.

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

[8] 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.

[9] K. Ma, H. Fu, T. Liu, Z. Wang, and D. Tao, "Deep blur mapping: Exploiting high-level semantics by deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5155–5166, 2018.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[11] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[16] U. Fang, J. Li, X. Lu, L. Gao, M. Ali, and Y. Xiang, "Self-supervised cross-iterative clustering for unlabeled plant disease images," *Neurocomputing*, vol. 456, pp. 36–48, 2021.

[17] X. Xiao, Z. Jin, Y. Hui, Y. Xu, and W. Shao, "Hybrid spatial–temporal graph convolutional networks for on-street parking availability prediction," *Remote Sensing*, vol. 13, no. 16, p. 3338, 2021.

[18] W. Shao, A. Prabowo, S. Zhao, P. Koniusz, and F. D. Salim, "Predicting flight delay with spatio-temporal trajectory convolutional network and airport situational awareness map," *arXiv preprint arXiv:2105.08969*, 2021.

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

[20] R. Huang, W. Feng, M. Fan, L. Wan, and J. Sun, "Multiscale blur detection by learning discriminative deep features," *Neurocomputing*, vol. 285, pp. 154–166, 2018.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[23] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, p. 107404, 2020.

[24] G. Xue, M. Zhong, J. Li, J. Chen, C. Zhai, and R. Kong, "Dynamic network embedding survey," *arXiv preprint arXiv:2103.15447*, 2021.

[25] Z. Li, X. Wang, J. Li, and Q. Zhang, "Deep attributed network representation learning of complex coupling and interaction," *Knowledge-Based Systems*, vol. 212, p. 106618, 2021.

[26] J. Chen, M. Zhong, J. Li, D. Wang, T. Qian, and H. Tu, "Effective deep attributed network representation learning with topology adapted smoothing," *IEEE Transactions on Cybernetics*, 2021.

[27] 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.

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

[29] X. Xiao, F. Yang, and A. Sadovnik, "Msdu-net: A multi-scale dilated u-net for blur detection," *Sensors*, vol. 21, no. 5, p. 1873, 2021.

[30] 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.

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