BGF-YOLO: Enhanced YOLOv8 with Multiscale Attentional Feature Fusion for Brain Tumor Detection

Ming Kang, Chee-Ming Ting(✉), Fung Fung Ting, and Raphaël C.-W. Phan

School of Information Technology, Monash University, Malaysia Campus, Subang Jaya, Malaysia
ting.cheeming@monash.edu

Abstract. You Only Look Once (YOLO)-based object detectors have shown remarkable accuracy for automated brain tumor detection. In this paper, we develop a novel BGF-YOLO architecture by incorporating Bi-level Routing Attention (BRA), Generalized feature pyramid networks (GFPN), and Fourth detecting head into YOLOv8. BGF-YOLO contains an attention mechanism to focus more on important features, and feature pyramid networks to enrich feature representation by merging high-level semantic features with spatial details. Furthermore, we investigate the effect of different attention mechanisms and feature fusions, detection head architectures on brain tumor detection accuracy. Experimental results show that BGF-YOLO gives a 4.7% absolute increase of mAP_{50} compared to YOLOv8x, and achieves state-of-the-art on the brain tumor detection dataset Br35H. The code is available at https://github.com/mkang315/BGF-YOLO.

Keywords: Medical image detection · lesion detection · YOLO · feature fusion · attention mechanism

1 Introduction

Detecting brain tumors in their early stages can lead to more effective treatments and better prognoses. Therefore, brain tumor detection is a critical aspect of medical diagnostics. Magnetic Resonance Imaging (MRI) is the best imaging test to visualize the brain and detect tumors. The You Only Look Once (YOLO) series has been demonstrated to detect brain tumors accurately. Kang et al. [9] proposed RCS-YOLO—a novel YOLO architecture with reparameterized convolution based on channel shuffle—on brain tumor detection and achieved a balance between accuracy and speed.

The YOLOv8 architecture [8] is mainly composed of the backbone and head parts, in which the neck is included in the head part. The backbone part, which is used for feature extraction, contains Conv, C2f (shortcut), and Spatial Pyramid Pooling Fast (SPPF) modules. The Conv, that is ConvBiSiLU (or CBS), and SPPF are the same as those in the YOLOv5 [7] architecture [27], where Conv is used to perform convolution operation on the input images and assist

C2f (shortcut) in feature extraction and SPPF enables adaptive-sized output. The C2f (shortcut) module is a lightweight convolutional structure compared to the C3 module in YOLOv5. Thus, the gradient flow of the model is enriched by connecting more branches across layers. Therefore, the more vital feature representation ability is enabled. The C2f (shortcut) module enhances the ability to express features through dense and residual structures, which changes the number of channels through split and concatenate operations according to scaling coefficients to reduce computational complexity and model capacity. The SPPF module at the end of the backbone part increases the sensitivity and captures the feature information of different levels in the images. The structures of Feature Pyramid Networks (FPN) [15] and Path Aggregation Network (PANet) [16] are used for multiscale feature fusion in the neck part. The FPN-PANet structure and C2f (without shortcut) modules fuse feature maps of different scales from the three stages of the backbone, aggregating shallow information to deep features. The head part employs a decoupled-head structure with a classification and regression (i.e., localization) prediction end to alleviate the conflict between classification and regression tasks, and an anchor-free mechanism to improve the detection of objects with irregular height and width. For bounding box classification, YOLOv8 employs binary cross-entropy loss while varifocal loss [32] is an alternative option. It can better handle the category imbalance situation and improve detection accuracy. For bounding box regression, YOLOv8 employs distribution focal loss [11,12,13] to overcome the problem of category imbalance and background category, allowing the network to quickly focus on the distribution of locations close to the object. It also uses the Complete Intersection over Union (CIoU) loss function [36] to alleviate the overlap between predicted and ground truth boxes. Nevertheless, some improvements potentially boost detection performance in the context of object detection in specific domains, which can still pose challenges using vanilla YOLOv8.

Recent improvements of YOLOv8 focus on attention mechanisms, multiscale feature fusion networks, and regression loss. Multi-Head Self-Attention mechanism was employed in MHSA-YOLOv8 [14]. A lightweight YOLOv8 [30] was proposed by combining a dual-path gated attention and feature enhancement module with the original YOLOv8s. An improved YOLOv8 with the neck structure of Asymptotic Feature Pyramid Network (AFPN) [31] was proposed in [5]. UAV-YOLOv8 [23] utilized the BiFormer block [37], focal fasternet blocks, and Wise-IoU (WIoU) [20] within the YOLOv8. Another improved YOLOv8 [35] also added Biformer in the backbone of YOLOv8 for insulator fault detection. DCA-YOLOv8 [29] employed deformable convolution and Coordinate Attention (CA) [3] within YOLOv8 for fast cattle detection. CSS-YOLO [17] respectively introduced the Swin Transformer and Convolution Block Attention Module (CBAM) [26] into YOLOv8's backbone and neck.

In this paper, we propose a novel model called BGF-YOLO, which enhances the detection performance of YOLOv8 by incorporating Bi-level Routing Attention (BRA) [37], Generalized-FPN (GFPN) [6], and Fourth detecting head. The contributions of this work are summarized as follows: 1) We reconstruct the

original neck part of YOLOv8 with a structured feature fusion network based on GFPN to facilitate effective feature fusion at different levels. 2) We leverage BRA for both dynamic and sparse attention mechanisms to focus on more salient features and reduce feature redundancy. 3) We add a fourth detecting head to enrich the scales of anchor boxes and optimize regression loss for detection. 4) To our best knowledge, this is the first use of enhanced YOLOv8 for brain tumor detection. The proposed modifications significantly improve tumor detection compared to YOLOv8. We also assess the effects of using various attention mechanisms, feature pyramid networks, and regression losses on the detection performance.

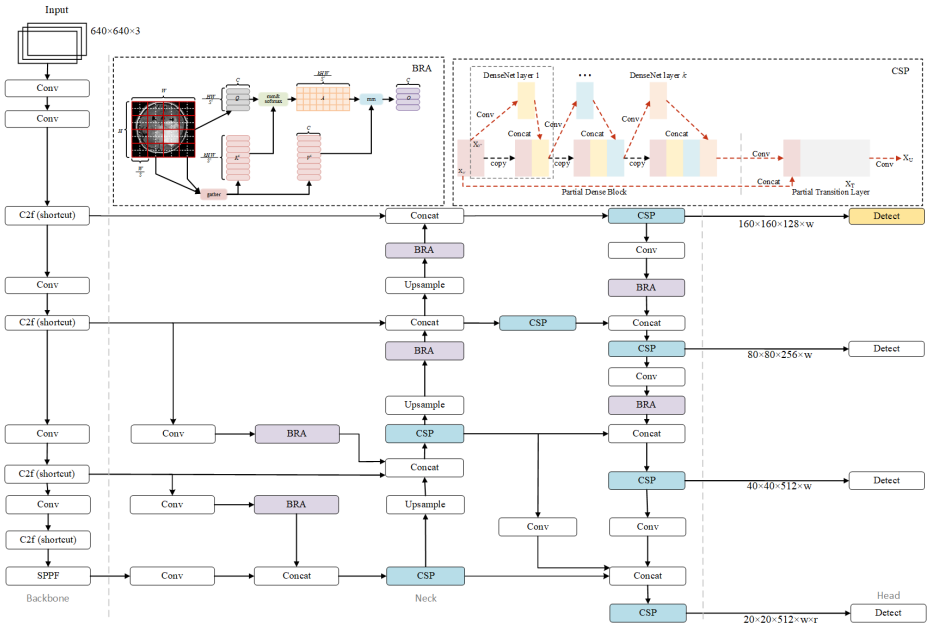

Fig. 1. Overview of BGF-YOLO. The architecture of BGF-YOLO is based on YOLOv8 and incorporates new modules colored Bi-level Routing Attention (BRA) [37], Cross Stage Partial DenseNet (CSP) [21], and enhanced detecting head. Conv, C2f (shortcut), SPPF, Concat, Upsample, and Detect are existing modules in the original YOLOv8 architecture [10].

2 Methods

Fig. 1 illustrates the architecture of the proposed BGF-YOLO. BGF-YOLO is characterized by a very deep and heavy neck, in contrast to the lightweight neck & head part in the YOLOv8. The backbone and head parts of the proposed

network are based on those of YOLOv8. Details of each part of the BGF-YOLO network structure are described in this section.

2.1 Enhanced GFPN for Multilevel Feature Fusion

FPN was first introduced to address the hierarchical feature fusion issue of convolutional neural networks (CNNs) and has been proven to effectively enhance the capabilities of deep learning models in handling object detection tasks, especially for detecting objects at various scales. PANet has been used to strengthen feature propagation and encourage information reuse, thus improving the feature pyramid's representational power. Bidirectional FPN (BiFPN) [19] adds bottom-up pathways to FPN with only a top-down pathway, resulting in bidirectional cross-scale connections to harness multiscale features efficiently. Generalized-FPN (GFPN) [6] employs the structure of dense link and queen fusion to produce better-fused features and uses the concat operation instead of the sum to perform feature fusion to reduce loss of information. AFPN uses an adaptive spatial fusion in an asymptotic process, which initially fuses two low-level features, then higher-level features, and finally top-level features, to enhance the significance of pivotal levels and mitigate the impact of contradictory information from different objects.

FPN and PANet were later used for multiscale feature fusion in the necks of YOLOv5 and YOLOv8. The difference in the neck parts between YOLOv5 and YOLOv8 is that the C2f (without shortcut) modules in YOLOv8 replace the C3 modules in YOLOv5 at the upsampling stage. FPN first extracts the feature maps in the CNNs and then uses upper sampling and coarse-grained feature maps to achieve the fusion of the feature maps in a top-down manner. In contrast, PANet fuses feature maps from bottom to top to ensure spatial information is accurately preserved. However, the combination of FPN and PANet can only support top-down and bottom-up feature fusion. The structures of BiFPN, AFPN, and GFPN are supposed to quickly integrate features into various levels and improve the effect of feature fusion by adding more levels to meet the needs of feature fusion at different levels.

We modify the structure of the FPN-PANet in the YOLOv8 to achieve multilevel feature fusion among different layers by strengthening the multipath fusion of the networks. Inspired by the GFPN and reparameterized GFPN-based DAMO-YOLO [28], we utilize Cross Stage Partial DenseNet (CSP) [21] to add skip connections and simultaneously share dense information across various spatial scales and non-adjacent levels of latent semantics by replacing C2f (without shortcut) and combining with Conv. This allows the model to handle both high-level semantic information and low-level spatial information with equal importance in the neck part. Brain tumor images can exhibit variability in terms of tumor size, shape, and location. Cross-stage connections allow the network to better adapt to this variability by incorporating features from different scales, improving the model's robustness to diverse tumor characteristics.

2.2 BRA-based Attentional Feature Fusion

The main idea of multiscale feature fusion networks in the neck part is to fuse feature maps extracted from different network layers to improve object detection performance at multiple scales. However, the feature fusion layer in YOLOv8 still suffers from the problem of redundant information from different feature maps. To overcome this limitation, we consider incorporating an attention mechanism to the feature fusion process in the YOLOv8 model.

The attention mechanism was originally proposed to weigh the importance of specific features relative to others. In the context of computer vision, there are five attention mechanisms that have great potential in improving the performance of object detection: Squeeze-and-Excitation (SE) [4], CBAM, Efficient Channel Attention (ECA) [22], CA, Receptive-Field Attention (RFA) [34], and BRA. The differences among them are that the SE and ECA belong to channel attention, the RFA and BRA handle spatial attention, and the CBAM and CA facilitate channel and spatial attention. SE is to adaptively recalibrate channel-wise feature responses by explicitly modeling the interdependencies between the channels of convolutional features. ECA only captures the local channel interdependencies without relying on global statistics to reduce computational requirements. The advantage of RFA is that it provides effective attention weight to realize convolutional kernel parameter sharing. BRA is a dynamic, query-aware sparse attention mechanism that enables a small subset of the most relevant key/value tokens for each query in a content-aware manner.

We improve the proposed GFPN-based feature fusion structure by adopting the BRA attention module to achieve effective multilevel feature fusion while avoiding redundant information across feature maps. The dynamic sparse attention can reduce redundant feature information and improve the model's detection accuracy, by applying the weight distribution of each channel and spatial position when integrating feature maps of different scales. We place the BRA modules behind the Conv or Upsample module in the feature fusion process to make the model only focus on specific areas after feature extraction. To further avoid information loss, skip connections in CSP modules enable the knowledge of the underlying feature maps to be reused in subsequent layers. BRA aims to eliminate a majority of non-pertinent key-value pairs input at a broader regional level, leaving only a select few relevant areas. Taking a feature map as input, the BRA first segments it into various areas and derives the query, key, and value through a linear transformation. The region-level relationship of queries and keys is entered into an adjacency matrix to construct a directed graph and pinpoint the association of specific key-value pairs. This essentially identifies which areas should be involved with each designated region. Lastly, multi-head self-attention is executed between individual tokens by utilizing the region-to-region routing index matrix. Through the bi-level routing optimization for multi-head self-attention, more attention is paid to the brain tumor part of the feature map, thereby improving the model's ability to detect brain tumors. The BRA attention module can be integrated into feature fusion to allow the model to focus more on relevant regions of the MRI scans where tumors are likely to be present. This

selective focus helps in capturing intricate details and subtle features associated with tumors.

This proposed method only uses the attention module BRA of BiFormer, which is different from the existing works [23,35] that add BiFormer into YOLOv8.

2.3 An Enhanced Detecting Head

The original YOLOv8 has three detecting heads with respective dimensions in height and width of 20×20, 40×40, and 80×80. In contrast, these heads still cannot meet the detection needs of brain tumor detection scenarios, which leads to the unsatisfactory detection accuracy of the model for larger objects than the original scales.

We introduce an additional 160×160 detecting head in the head part aligned with the new structure of feature fusion networks in the neck part to improve the detection capacity for objects in various scales. The new scale-detecting head is added as the fourth detecting head next to the original 80×80 detection scales of YOLOv8. It fuses the shallow information of the first C2f (shortcut) module from the input images, incorporating additional feature fusion networks. The one more prediction head we add enhances the model to detect objects in richer scales.

The additional detecting head can classify and localize different brain tumor sizes within the model. In the case of brain tumor detection, this can be beneficial when there are different sizes of tumors with distinct characteristics. Extensive layers of detection heads can facilitate progressive detection, where the model first identifies potential regions of interest and then refines its predictions through subsequent processing stages. This step-wise approach can lead to more accurate and robust tumor detection.

3 Experiments and Results

3.1 Dataset Details

We evaluated the performance of the proposed BGF-YOLOv8 on the public brain tumor image dataset Br35H [2] which contains 801 MRI images with annotated brain tumors. The dataset was divided into the train set of 500 images, the validation set of 201 images, and the test set of 100 images. All the results are tested on the test set.

3.2 Implementation Details

The BGF-YOLO was trained and tested with Intel® Xeon® Platinum 8255C CPU @ 2.50GHz and NVIDIA® GeForce GTX® 1060 6GB GPU. We implemented the proposed methods based on YOLOv8 extra large version (YOLOv8x). The hyperparameters used in the training of BGF-YOLO and other comparison methods are the same as YOLOv8x. The training parameter batch size is set to

5, and the epoch is 120 at the training stage. The optimizer uses the stochastic gradient descent with the initial and final learning rate of 0.01 and momentum of 0.937.

3.3 Results

For a fair comparison, we choose the version with the best performance of the competing models and use the same evaluation metrics as those used to evaluate them. As shown in Table 1, BGF-YOLO achieved 1.2%, 4.7% and 0.7% absolute increase in precision, mean average precision mAP_{50} and $mAP_{50:95}$ respectively, compared to YOLOv8x. It also outperforms YOLOv9-E [24], YOLOv10-X [25], RCS-YOLO, and DAMO-YOLO-L*,. BGF-YOLO surpasses not only the baseline YOLOv8 model but also a GFPN-neck detector DAMO-YOLO and a high-accuracy and fast detector RCS-YOLO. As shown in supplementary material Section 1, in qualitative comparison, brain tumor tumors are more accurately detected by BGF-YOLO than YOLOv8.

Table 1. Performance comparison of YOLOv8x, YOLOv9-E, YOLOv10-X, DAMO-YOLO-L*, RCS-YOLO, and our proposed BGF-YOLO. * in DAMO-YOLO-L* indicates distillation was employed. The original code of all DAMO-YOLO versions only prints average precision and average recall. The best results are shown in bold.

Model	Precision	Recall	mAP_{50}	$mAP_{50:95}$
YOLOv8x [8]	0.907	0.881	0.927	0.646
YOLOv9-E [24]	0.927	0.869	0.919	0.630
YOLOv10-X [25]	0.916	0.808	0.880	0.603
RCS-YOLO [9]	0.908	0.885	0.878	0.580
DAMO-YOLO-L* [28]	–	–	0.900	0.610
BGF-YOLO (Ours)	**0.919**	**0.926**	**0.974**	**0.653**

3.4 Ablation Study

We conducted a series of ablation studies to assess the advantages of incorporating each method in the proposed BGF-YOLO model, and to investigate the effect of using different techniques in each method on the detection performance through the following extensive experiments. See supplementary material Section 2: Tables of Ablation Study.

Ablation Study on Overall Architecture We evaluated four incomplete BGF-YOLO models by removing each method respectively. Supplementary material Table 1 shows that BRA, GFPN, the fourth head, and GIoU all contribute to the accuracy improvement of BGF-YOLO. The w/o GFPN means using the original neck structure FPN-PANet of YOLOv8. Adding the fourth detection head gives the most impact on the overall accuracy improvement especially for mAP_{50}, followed by GFPN and BRA.

Effect of Different Multiscale Feature Fusion Structures We compared the proposed BGF-YOLO with BBF-YOLO and BAF-YOLO, which respectively replace GFPN with BiFPN and AFPN in the neck part of BGF-YOLO for feature fusion. As shown in Supplementary material Table 2, the precision, mAP_{50}, and $mAP_{50:95}$, expect recall of BGF-YOLO with GFPN structure are much higher than the model with BiFPN and AFPN structures.

Effect of Different Attention Mechanisms We investigated different attention mechanisms with the proposed BGF-YOLO model. The first letters of the model names listed in supplementary material Table 3 represent the attention mechanisms used, which means S, E, C, A, R, and B denote SE, ECA, CBAM, CA, RFA, and BRA. BRA gives the largest performance improvement among the different attention mechanisms compared to the other five alternative ones. Meanwhile, CBAM (i.e., CGF-YOLO) ranks second behind BRA (i.e., BGF-YOLO) in terms of mAP_{50} and has higher values in precision than BRA. Although the $mAP_{50:95}$ of ECA (i.e., EGF-YOLO) and CA (i.e., AGF-YOLO) are higher than that of BRA, the mAP_{50} of ECA and CA are much lower than that of BRA.

Effect of Different Regression Losses We performed an ablation study on the influence of regression losses, including Generalized IoU (GIoU) [18] where the distance between two axis-aligned rectangles is calculated, Distance-IoU (DIoU) [36] optimal objectives of which are less than CIoU, Efficient IoU (EIoU) [33] which explicitly measures the discrepancies of three geometric factors, Scylla-IoU (SIoU) [1] where penalty metrics are redefined, WIoU v3 which is a two-layer attention-based with a dynamic nonmonotonic focusing mechanism regression loss function. These loss functions are represented by C, E, S, and W as the fourth letters of model names in supplementary material Table 4. Compared to other regression losses, The original regression loss CIoU in YOLOv8 has better robustness of the bounding box for object detection. The mAP_{50} of DIoU (i.e., BGFD-YOLO) is close to that of CIoU (i.e., BGF-YOLO), which indicates DIoU is a competitor to CIoU. In terms of $mAP_{50:95}$, those of GIoU (i.e., BGFG-YOLO) and EIoU (i.e., BGFE-YOLO) are higher than that of CIoU. Which regression loss is a better choice depends on the criterion of the specific scenario. In this case, we choose mAP_{50} as the main metric for brain tumor detection, and therefore, CIoU is selected as regression loss in the proposed BGF-YOLO.

4 Conclusion

We developed a novel BGF-YOLO model building on YOLOv8 for accurate detection of brain tumors from MRI. We show that the object detection capability of YOLOv8 is substantially enhanced by the optimization of the GFPN feature fusion structure, BRA attention mechanism, and adding a detecting head in the

BGF-YOLO model. These modifications enable weighted feature fusion at different levels and at richer scales and produce high-quality anchor boxes with dynamic focusing mechanisms. Besides, the proposed modules in BGF-YOLO are better than the other alternative techniques, as shown in a series of experimental evaluations on different feature fusion structures, attention mechanisms, and regression losses. Our proposed BGF-YOLO becomes the current state-of-the-art model on the brain tumor detector dataset Br35H.

Acknowledgement. This work was supported by the Monash University Malaysia and the Ministry of Higher Education, Malaysia under Fundamental Research Grant Scheme FRGS/1/2023/ICT02/MUSM/02/1.

Disclosure of Interests. The author has no conflict of interest related to the article.

References

1. Gevorgyan, Z.: SIoU loss: more powerful learning for bounding box regression. arXiv preprint arXiv:2205.12740 (2022)
2. Hamada, A.: Br35H :: brain tumor detection 2020. Kaggle (2020). https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection
3. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13708–13717. IEEE, Piscataway (2021)
4. He, K., Zhang, X., Ren, S., Sun, J.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141. IEEE, Piscataway (2018)
5. Huang, Z., Li, L., Krizek, G.C., Sun, L.: Research on traffic sign detection based on improved YOLOv8. J. Comput. Commun. **11**(7), 226–232 (2023)
6. Jiang, Y., Tan, Z., Wang, J., Sun, X., Lin, M., Li, H.: GiraffeDet: a heavy-neck paradigm for object detection. In: 2022 International Conference on Learning Representations (ICLR). (2022)
7. Jocher, G.: YOLO by ultralytics (version 5.7.0). GitHub (2022). https://github.com/ultralytics/yolov5
8. Jocher, G., Chaurasia, A., Qiu, J.: YOLO by ultralytics (version 8.0.190). GitHub (2023). https://github.com/ultralytics/ultralytics
9. Kang, M., Ting, C.-M., Ting, F.F., Phan, R.C.-W.: RCS-YOLO: a fast and high-accuracy object detector for brain tumor detection. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14223, 600–610. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43901-8_57
10. King, R.: Brief summary of YOLOv8 model structure. GitHub (2023). https://github.com/ultralytics/ultralytics/issues/189
11. Li, X., et al.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems 33 (NeurIPS), pp. 21002–21012. Curran Associates, New York (2020)

12. Li, X., Wang, W., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss v2: learning reliable localization quality estimation for dense object detection. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11632–11641. IEEE, Piscataway (2021)
13. Li, X., Lv, C., Wang, W., Li, G., Yang, L., Yang, J.: Generalized focal loss: Towards efficient representation learning for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. **45**(3), 3139–3153 (2023)
14. Li, P., Zheng, J., Li, P., Long, H., Li, M., Gao, L.: Tomato maturity detection and counting model based on MHSA-YOLOv8. Sens. **23**(15), 6701 (2023)
15. Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2117–2125. IEEE, Piscataway (2017)
16. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8759–8768. IEEE, Piscataway (2018)
17. Lu, L.: Improved YOLOv8 detection algorithm in security inspection image. arXiv preprint arXiv:2308.06452 (2023)
18. Rezatofighi, H., Tsoi, J., Gwak, J., Sadeghian, A., Reid, I.: Generalized intersection over union: a metric and a loss for bounding box regression. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 658–666. IEEE, Piscataway (2019)
19. Tan, M., Pang, R., Le, Q.V.: EfficientDet: scalable and efficient object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10781–10790. IEEE, Piscataway (2020)
20. Tong, Z., Chen, Y., Xu, Z., Yu, R.: Wise-IoU: bounding box regression loss with dynamic focusing mechanism. arXiv preprint arXiv:2301.10051 (2023)
21. Wang, C.-Y., Liao, H.-Y.M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., Yeh, I.-H.: CSPNet: a new backbone that can enhance learning capability of CNN. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1571–1580. IEEE, Piscataway (2020)
22. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-Net: efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11534–11542. IEEE, Piscataway (2020)
23. Wang, G., Chen, Y., An, P., Hong, H., Hu, J., Huang, T.: UAV-YOLOv8: a small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios. Sens. **23**(16), 7190 (2023)
24. Wang, C.-Y., Yeh, I.-H., Liao, H.-Y.M.: YOLOv9: learning what you want to learn using programmable gradient information. arXiv preprint arXiv:2402.13616 (2024)
25. Wang, A., et al.: YOLOv10: real-time end-to-end object detection. arXiv preprint arXiv:2405.14458 (2024)
26. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
27. Wu, Z.: YOLOv5 (6.0/6.1) brief summary. GitHub (2022). https://github.com/ultralytics/yolov5/issues/6998
28. Xu, X., Chen, Y., Xu, Z., Yu, R.: DAMO-YOLO: a report on real-time object detection design. arXiv preprint arXiv:2211.15444 (2023)

29. Yang, W., et al.: Deformable convolution and coordinate attention for fast cattle detection. Comput. Electron. Agric. **211**, 108006 (2023)
30. Yang, G., Wang, J., Nie, Z., Yang, H., Yu, S.: A lightweight YOLOv8 tomato detection algorithm combining feature enhancement and attention. Agron. **13**(7), 1824 (2023)
31. Yang, G., Lei, J., Zhu, Z., Cheng, S., Feng, Z., Liang, R.: AFPN: asymptotic feature pyramid network for object detection. arXiv preprint arXiv:2306.15988 (2023)
32. Zhang, H., Wang, Y., Dayoub, F., Sünderhauf, N.: VarifocalNet: an IoU-aware dense object detector. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8510–8519. IEEE, Piscataway (2021)
33. Zhang, Y.-F., Ren, W., Zhang, Z., Jia, Z., Wang, L., Tan, T.: Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing **506**, 146–157 (2022)
34. Zhang, Y., et al.: RFAConv: innovating spatial attention and standard convolutional operation. arXiv preprint arXiv:2304.03198 (2023)
35. Zhang, Y., Wu, Z., Wang, X., Fu, W., Ma, J., Wang, G.: Improved YOLOv8 insulator fault detection algorithm based on BiFormer. In: 2023 IEEE 5th International Conference on Power, Intelligent Computing and Systems (ICPICS), pp. 962–965. IEEE, Piscataway (2023)
36. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-IoU loss: faster and better learning for bounding box regression. Proc. AAAI Conf. Artif. Intell., **34**(07), 12993–13000 (2020)
37. Zhu, L., Wang, X., Ke, Z., Zhang, W., Lau, R.W.H.: BiFormer: vision transformer with bi-Level routing attention. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10323–10333. IEEE, Piscataway (2023)