Transformer Framework and YOLO Framework for Object Detection

Subjects: Computer Science, Artificial Intelligence

Contributor: Lingtong Min , Ziman Fan , Qinyi Lv , Mohamed Reda , Linghao Shen , Binglu Wang

Object detection for remote sensing is a fundamental task in image processing of remote sensing; as one of the core components, small or tiny object detection plays an important role.

small object detection remote sensing images transformer YOLO

1. Introduction

Remote sensing object detection is a prominent and consequential application within the realm of remote sensing image processing [1]. It aims to accurately identify and locate specific target instances within an image. Within this domain, remote sensing small object detection holds particular importance as it focuses on detecting objects in remote sensing images that occupy a very small area or consist of only a few pixels. Detecting small objects is considerably more challenging than detecting larger objects, resulting in lower accuracy rates [2]. In recent years, small object detection based on convolutional neural networks (CNNs) has rapidly developed with the rapid growth of deep learning [3]. Small object detection often faces challenges such as limited information on small objects, scarcity of positive samples, and imbalanced classification. To tackle this challenge, researchers and experts have put forth diverse deep neural network methodologies, encompassing CNNs, GANs, RNNs, and transformers, to tackle the detection of small objects, encompassing remote sensing small objects. To improve the detection of small objects, Liu W et al., proposed the YOLOV5-Tassel network, which introduced the SimAM module in front of each detection head to extract the features of interest [4]. Li J. et al., suggested using GAN models to generate high-resolution images of small objects, narrowing the gap between small and large objects, and improving the detection capability of tiny objects [5]. Xu W et al. integrated contextual information into the Swin Transformer and designed an advanced framework called the foreground-enhanced attention Swin Transformer (FEA-Swin) [6]. Although the accuracy of detecting small objects has improved, the speed has been somewhat compromised. Zhu X. et al., proposed the YOLOv5-THP model, which is based on YOLOv5 and adds a transformer model with an attention mechanism to the detection head [7]. While this enhances the network's performance in detecting small objects, it also brings a significant computing burden.

In the field of remote sensing, detecting small objects remains challenging due to large image scales, complex and varied backgrounds, and unique shooting perspectives. Cheng et al. propose a model training regularization method that enhances the performance of detection of small or tiny objects in remote sensing by exploiting and incorporating global contextual cues and image-level contextual information [8]. Liu J. et al., added a dilated convolution module to the FPN and designed a relationship connection attention module to automatically select and refine features, combining global and local attention to achieve the detection task of small objects in remote sensing [9]. Cheng et al., proposed an end-to-end cross-scale feature fusion (CSFF) framework based on the feature pyramid network (FPN), which inserted squeeze-and-excitation (SE) modules at the top layer to achieve better detection of tiny objects in optical remote sensing images [10]. Dong et al., proposed a CNN method based on balanced multi-scale fusion (BMF-CNN), which fused high- and low-level semantic information to improve the detection performance of tiny objects in remote sensing [11]. Liang X. et al., proposed a single-shot detector (FS-SSD) based on feature fusion and scaling to better adapt to the detection of tiny or small objects in remote sensing. FS-SSD added a scaling branch in the deconvolution module and used two feature pyramids generated by the deconvolution module and feature fusion module together for prediction, improving the accuracy of object detection [12]. Xu et al., designed a transformer-guided multi-interaction network (TransMIN) using local–global feature interaction (LGFI) and cross-view feature interaction (CVFI) modules to enhance the performance of small object detection in remote sensing. However, this improvement unavoidably introduces a computational burden [13]. Li et al., proposed a transformer that aggregates multi-scale global spatial positions to enhance small object detection performance but it also comes with a computational burden [14]. To reduce the computational cost of the transformer, Xu et al., improved the lightweight Swin transformer and designed a Local Perception Swin transformer (LPSW) backbone network to enhance small-scale detection accuracy [15]. Gong et al., designed an SPH-YOLOv5 model based on Swin Transformer Prediction Heads (SPHs) to balance the accuracy and speed of small object detection in remote sensing [16]. Although many experts and scholars are studying the balance between detection accuracy and inference speed, achieving an elegant balance remains a challenging problem [17][18][19][20][21].

Considerable advancements have been achieved in the utilization of transformers [6][7][13][14][15][16] for small object detection within the remote sensing domain. The exceptional performance of the Contextual Transformer (CoT) [22] in harnessing spatial contextual information, thereby offering a fresh outlook on transformer design, merits significant attention. In the domain of remote sensing, small target pixels are characterized by a scarcity of spatial information but a profusion of channel-based data. Consequently, the amalgamation and modeling of spatial and channel information assume paramount importance. Furthermore, transformers impose notable demands on computational resources and network capacity, presenting a challenge in striking an optimal balance between detection accuracy and processing speed for small object detection in the remote sensing discipline. Meanwhile, Bar M et al. demonstrated that the background is critical for human recognition of objects [18]. Empirical research in computer vision has also shown that both traditional methods [19] and deep learning-based methods [12] can enhance algorithm performance by properly modeling spatial context. Moreover, He K. et al., have proven that residual structures are advantageous for improving network performance [17][20]. Finally, researchers note that the classification and regression tasks of object detection focus on the salient features and boundary features of the target, respectively [23]. Therefore, a decoupled detection head incorporating residual structure as well as channel and spatial context knowledge should have a positive impact on the detection of small or tiny objects.

2. Transformer Framework for Object Detection

The transformer structure, based on self-attention, first appeared in NLP tasks. Compared to modern convolutional neural networks (CNN) [24], the Vision Transformer has made impressive progress in the field of computer vision. After Dosovitskiy A et al. successfully introduced transformers into computer vision [25], many scholars turned to transformers [26][27][28]. In object detection, DETR [29] and Pix2seq [30] are the earliest transformer detectors that define two different object detection paradigms. However, transformers have many parameters, require high computing power and hardware, and are not easily applicable. To apply transformers on mobile devices, Mehta S et al. proposed a lightweight MobileVIT series [31][32][33], which achieved a good balance between accuracy and real-time performance, and has been widely used in risk detection [34], medicine [35], and other fields. A major advantage of transformers is that they can use the attention mechanism to model the global dependence of input data, obtain longer-term global information, and ignore the connection between local contexts. To address this problem, Li Y. et al., proposed a lightweight CoT [22] self-attention module to capture contextual background information on 2D feature maps. It can extract information between local contexts while capturing global dependencies for more adequate information exchange. In this research, researchers use CoT to exploit the global characteristics of spatial context and channels. Based on the original structure, researchers added the global residual and local fusion structures to further tap and utilize the characteristics of space and channels.

3. YOLO Framework for Object Detection

In 2015, YOLO [36] introduced a one-stage object detection method that combined candidate frame extraction, CNN feature learning, and NMS optimization to simplify the network structure. The detection speed was nearly 10 times faster than R-CNN, making real-time object detection possible with the computing power available at that time. However, it was not suitable for detecting small objects. YOLOv2 [37] added optimization strategies such as batch normalization and a dimensional clustering anchor box based on v1 to improve the accuracy of object regression and positioning. YOLOv3 [38] added the residual structure and FPN structure based on v2 to further improve the detection performance of small objects. The network framework structure after YOLOv3 can be roughly divided into three parts, backbone, neck, and head. Subsequent versions have optimized internal details to varying degrees. For example, YOLOv4 [39], based on v3, further optimized the backbone network and activation function, and used Mosaic data enhancement to improve the robustness and reliability of the network. YOLOv5 [40] added the focus structure based on v4 and accelerated the training speed by slicing. YOLOv6 [41] introduced RepVGG in the backbone, proposed a more efficient EfficientRep block, and simplified the design of the decoupling detection head to improve the detection efficiency. YOLOv7 [42] adopted the E-ELAN structure in the neck part, which reduces the inference speed, and used the auxiliary head training method. At present, YOLOv7 is one of the more advanced object detection networks due to its real-time characteristics. It is widely used in fields with high time requirements such as industrial equipment inspection [43], sea rescue [44], and aquaculture [45]. Therefore, researchers use YOLOv7, one of the powerful benchmarks, as the benchmark model.

4. Detection Head Framework for Object Detection

In the object detection task, there are two tasks: classification and regression, which respectively output the classification and bounding box position of the object. Song G. et al., pointed out that the focus of the classification and regression tasks is different [23]. Specifically, classification pays more attention to the texture content of the object, while regression pays more attention to the edge information of the object. Wu Y et al. suggested that it may be better to divide classification and regression tasks into FC-head and Conv-head [46]. In the single-stage model, YOLOX [47] adopts the decoupling head structure that separates the classification and regression branches and adds two additional 3 × 3 convolutional layers. This improves detection accuracy at the cost of inference speed. Building upon this approach, YOLOv6 takes into account the balance between the representation ability of related operators and the hardware computing overhead and adopts the Hybrid Channels strategy to redesign a more efficient decoupling head structure that reduces the cost while maintaining accuracy. They also mitigate the additional latency overhead of 3 × 3 convolutions in the decoupled detection head. Feng C. et al., use feature extractors to learn task interaction features from multiple convolutional layers to enhance the interaction between classification and localization [48]. They also pointed out that the interaction characteristics of different tasks may vary due to the classification and localization goals. To resolve the feature conflict introduced between the two tasks, they designed a layer attention mechanism that focuses on different types of features such as different layers and receptive fields. This mechanism helps to resolve a certain degree of feature conflict between the two tasks.

References

1. Wang, B.; Zhao, Y.; Li, X. Multiple instance graph learning for weakly supervised remote sensing object detection. IEEE Trans. Geosci. Remote Sens. 2021, 60, 5613112.
2. Tong, K.; Wu, Y. Deep learning-based detection from the perspective of tiny objects: A survey. Image Vis. Comput. 2022, 123, 104471.
3. Wu, Y.; Zhang, K.; Wang, J.; Wang, Y.; Wang, Q.; Li, Q. CDD-Net: A context-driven detection network for multiclass object detection. IEEE Geosci. Remote Sens. Lett. 2020, 19, 8004905.
4. Liu, W.; Quijano, K.; Crawford, M.M. YOLOv5-Tassel: Detecting tassels in RGB UAV imagery with improved YOLOv5 is based on transfer learning. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2022, 15, 8085–8094.
5. Li, J.; Liang, X.; Wei, Y.; Xu, T.; Feng, J.; Yan, S. Perceptual generative adversarial networks for small object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1222–1230.
6. Xu, W.; Zhang, C.; Wang, Q.; Dai, P. FEA-swin: Foreground enhancement attention swin transformer network for accurate UAV-based dense object detection. Sensors 2022, 22, 6993.
7. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the

IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.

8. Cheng, G.; Lang, C.; Wu, M.; Xie, X.; Yao, X.; Han, J. Feature enhancement network for object detection in optical remote sensing images. J. Remote Sens. 2021, 2021, 9805389.
9. Liu, J.; Yang, D.; Hu, F. Multiscale object detection in remote sensing images combined with multi-receptive-field features and relation-connected attention. Remote Sens. 2022, 14, 427.
10. Cheng, G.; Si, Y.; Hong, H.; Yao, X.; Guo, L. Cross-scale feature fusion for object detection in optical remote sensing images. IEEE Geosci. Remote Sens. Lett. 2020, 18, 431–435.
11. Dong, Z.; Lin, B. BMF-CNN: An object detection method based on multi-scale feature fusion in VHR remote sensing images. Remote Sens. Lett. 2020, 11, 215–224.
12. Liang, X.; Zhang, J.; Zhuo, L.; Li, Y.; Tian, Q. Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis. IEEE Trans. Circuits Syst. Video Technol. 2019, 30, 1758–1770.
13. Xu, G.; Song, T.; Sun, X.; Gao, C. TransMIN: Transformer-Guided Multi-Interaction Network for Remote Sensing Object Detection. IEEE Geosci. Remote Sens. Lett. 2022, 20, 6000505.
14. Li, Q.; Chen, Y.; Zeng, Y. Transformer with transfer CNN for remote-sensing-image object detection. Remote Sens. 2022, 14, 984.
15. Xu, X.; Feng, Z.; Cao, C.; Li, M.; Wu, J.; Wu, Z.; Shang, Y.; Ye, S. An improved swin transformer-based model for remote sensing object detection and instance segmentation. Remote Sens. 2021, 13, 4779.
16. Gong, H.; Mu, T.; Li, Q.; Dai, H.; Li, C.; He, Z.; Wang, W.; Han, F.; Tuniyazi, A.; Li, H.; et al. Swin-Transformer-Enabled YOLOv5 with Attention Mechanism for Small Object Detection on Satellite Images. Remote Sens. 2022, 14, 2861.
17. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
18. Bar, M. Visual objects in context. Nat. Rev. Neurosci. 2004, 5, 617–629.
19. Carbonetto, P.; De Freitas, N.; Barnard, K. A statistical model for general contextual object recognition. In Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic, 11–14 May 2004; pp. 350–362.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Part IV 14, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 630–645.

21. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; et al. Speed/accuracy trade-offs for modern convolutional object detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7310–7311.
22. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual transformer networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2022, 45, 1489–1500.
23. Song, G.; Liu, Y.; Wang, X. Revisiting the sibling head in object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11563–11572.
24. Liu, Y.; Zhang, Y.; Wang, Y.; Hou, F.; Yuan, J.; Tian, J.; Zhang, Y.; Shi, Z.; Fan, J.; He, Z. A survey of visual transformers. arXiv 2021, arXiv:2111.06091.
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020, arXiv:2010.11929.
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
27. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578.
28. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 558–567.
29. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Part I 16, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
30. Chen, T.; Saxena, S.; Li, L.; Fleet, D.J.; Hinton, G. Pix2seq: A language modeling framework for object detection. arXiv 2021, arXiv:2109.10852.
31. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. arXiv 2021, arXiv:2110.02178.
32. Mehta, S.; Rastegari, M. Separable self-attention for mobile vision transformers. arXiv 2022, arXiv:2206.02680.

33. Wadekar, S.N.; Chaurasia, A. Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features. arXiv 2022, arXiv:2209.15159.

34. Tong, H.; Peng, T.; Jiang, X. A Lightweight Risk Advertising Image Detection Method Based on Mobile-ViT. In Proceedings of the 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Dalian, China, 11–12 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1249–1253.

35. Marefat, A.; Joloudari, J.H.; Rastgarpour, M. A Transformer-Based Algorithm for Automatically Diagnosing Malaria Parasite in Thin Blood Smear Images Using MobileViT; Technical Report; EasyChair: Manchester, UK, 2022.

36. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.

37. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.

38. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. arXiv 2018, arXiv:1804.02767.

39. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. arXiv 2020, arXiv:2004.10934.

40. Yolov5. Available online: https://github.com/ultralytics/yolov5 (accessed on 15 March 2023).

41. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. arXiv 2022, arXiv:2209.02976.

42. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv 2022, arXiv:2207.02696.

43. Hussain, M.; Al-Aqrabi, H.; Munawar, M.; Hill, R.; Alsboui, T. Domain Feature Mapping with YOLOv7 for Automated Edge-Based Pallet Racking Inspections. Sensors 2022, 22, 6927.

44. Zhao, H.; Zhang, H.; Zhao, Y. Yolov7-sea: Object detection of maritime uav images based on improved yolov7. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 233–238.

45. Jiang, K.; Xie, T.; Yan, R.; Wen, X.; Li, D.; Jiang, H.; Jiang, N.; Feng, L.; Duan, X.; Wang, J. An Attention Mechanism-Improved YOLOv7 Object Detection Algorithm for Hemp Duck Count Estimation. Agriculture 2022, 12, 1659.

46. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking classification and localization for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10186–10195.

47. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. arXiv 2021, arXiv:2107.08430.
48. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. Tood: Task-aligned one-stage object detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE Computer Society: Piscataway, NJ, USA, 2021; pp. 3490–3499.

Retrieved from https://encyclopedia.pub/entry/history/show/109535