YOLO-SH:An Object Detection Model for Small Targets

Honggeng Zhang

***Abstract*—Achieving accurate object detection has always been a highly challenging task in the field of computer science, particularly in remote sensing image object detection. The possibility of objects appearing in any direction adds to the difficulty of achieving precise object detection. Additionally, the presence of small objects and the influence of environmental factors pose significant challenges for current deep learning-based object detection algorithms. In order to address these challenges, this paper proposes a target detection algorithm called YOLO-SH, which is based on YOLOv8. It improves the detection performance of objects appearing in any direction by incorporating the Swin Transformer structure into YOLO. The Swin Transformer utilizes sliding windows and hierarchical structures to achieve more efficient and flexible computation, and it can capture global information from multi-scale images with better feature extraction capabilities. Furthermore, to improve the detection performance for small objects, an additional prediction head is added, which effectively detects multi-scale objects based on the three original prediction heads in YOLOv8. Through extensive experiments on the DIOR dataset, our YOLO-SH model demonstrates excellent performance in remote sensing image detection.**

***Index Terms*—YOLOv8, DIOR, Swin Transformer.**

I. INTRODUCTION

Object detection is a task in the field of computer vision that aims to identify and localize objects of interest in digital images or videos. This task involves extracting features from the images and using models to classify and regress their positions, accurately determining the location and category of the target objects.

A remote sensing dataset is a collection of data used for remote sensing image processing and analysis. It includes image data acquired from satellites, aircraft, or other remote sensing sensors. These datasets are typically used for supervised learning, where the images are labeled with annotated regions containing the objects of interest, along with their corresponding class labels. Remote sensing datasets cover various application scenarios such as land use classification, object detection, and feature recognition. These datasets exhibit diverse characteristics such as different spatial resolutions, spectral information, and multi-temporal observations, to cater to the needs of various application domains.

For conducting object detection tasks, remote sensing datasets typically consist of a large number of annotated images with bounding boxes indicating the positions and boundaries of the target objects in the images. By utilizing these labeled data for training, object detection algorithms can learn to extract features from the images and accurately identify and locate objects. These datasets may also include various object categories such as buildings, roads, vegetation, etc., to cover a wide range of scenes and application requirements.

In object detection, remote sensing datasets pose some challenges that need to be addressed. Firstly, remote sensing images often have high resolution and large size, resulting in large-scale datasets that require significant computational and storage resources for processing and storage. This presents challenges for training and testing object detection algorithms.

Secondly, target objects in remote sensing images exhibit significant variations in terms of size, shape, pose, density, and occlusion. These variations make the object detection task more difficult as algorithms need to be capable of handling diverse targets. Detection performance may be limited for small-sized objects due to their limited pixel information and low contrast.

Furthermore, remote sensing images often contain noise, shadows, clouds, and other interferences, which can obscure or blur the visual features of the target objects, negatively impacting the accuracy of object detection algorithms. Hence, effective handling of these noises and interferences needs to be considered when processing remote sensing datasets to enhance the robustness of object detection.

Overall, remote sensing datasets in object detection present challenges related to their scale, diverse object variations, and potential noise and interference. These challenges necessitate the application of specialized techniques and methods to address them.

II. RELATED WORK

A. YOLO

YOLO[1] is a computer vision algorithm used for object detection. YOLO stands for You Only Look Once, meaning that it can perform object detection in a single pass. Compared to traditional object detection methods, YOLO offers higher real-time performance and efficiency. It achieves end-to-end object detection by simultaneously predicting the object's position and class in a single neural network. It utilizes a Convolutional Neural Network (CNN) architecture to extract image features and combines them with prediction layers to generate the final detection results. YOLO also incorporates the concept of anchor boxes, which are a predefined set of

Manuscript received February 04, 2024
Honggeng Zhang, School of computer science and technology, Tiangong University, Tianjin, China

bounding box sizes and ratios that help to better adapt to objects of different sizes and shapes. The main advantages of YOLO over traditional object detection algorithms are its speed and accuracy. Since YOLO only requires a single forward pass to obtain the positions and class probabilities of all detected objects, it enables real-time object detection suitable for video processing and real-time applications. Moreover, YOLO demonstrates excellent overall performance, achieving high object recognition accuracy while maintaining low false detection rates.

YOLOv8 is a computer vision algorithm for object detection, which stands for "You Only Look Once." It achieves real-time and efficient object detection by simultaneously predicting the positions and categories of objects in a single neural network. YOLOv8 uses a convolutional neural network (CNN) architecture to extract image features and combines them with prediction layers to produce the final detection results. It also employs anchor boxes, a predefined set of bounding box sizes and ratios, to better adapt to objects of different sizes and shapes. YOLOv8 excels in terms of speed and accuracy compared to traditional object detection methods. Due to its one-pass forward pass to predict the positions and categories of all objects, YOLOv8 enables real-time object detection suitable for video processing and real-time applications. Overall, YOLOv8 is an advanced object detection algorithm that combines deep learning and computer vision techniques to achieve fast and accurate object detection.

YOLOv8 has several advantages. First, it offers real-time performance by efficiently processing object detection, making it suitable for applications that require rapid response. Second, YOLOv8 achieves high accuracy in object detection. By utilizing deeper and more powerful CNN architectures, it effectively extracts rich and detailed image features, resulting in improved detection precision. Third, YOLOv8 handles objects of different scales through the introduction of multiscale feature maps, enhancing detection robustness and generalization ability. Lastly, YOLOv8 supports multitask detection of multiple object categories. Applying multitask training, it can simultaneously learn different classes of objects, improving overall detection performance.

However, there are some limitations to consider. YOLOv8 demands higher GPU resources due to its deep and complex network, making it challenging to train and infer on devices with limited computational power. Furthermore, YOLOv8 has a relatively large model size, requiring more storage space for model storage and deployment, which may restrict its usage in resource-constrained environments. Additionally, YOLOv8 exhibits relatively weaker detection performance for small objects since it relies on specific-scale feature maps for object detection, which may result in some missed detections.

B. Transformer

Transformer[2] is a neural network architecture based on self-attention mechanism used for processing sequential data, especially widely used in the field of Natural Language Processing. Unlike Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), Transformer does not rely on the sequential order of the input elements. Instead, it uses self-attention mechanism to capture non-linear dependencies and longer-range contextual information between elements effectively. A Transformer network consists of stacked encoders and decoders, each containing multi-head self-attention and feed-forward neural networks. Multi-head self-attention allows each element to attend to other elements in the sequence simultaneously and learn their relevancy, while the feed-forward neural network applies non-linear transformations to each element's features. During training, Transformer uses masked attention mechanism to dynamically hide subsequent elements in the target sequence, avoiding the model from relying on future information while generating sequences. In the field of Natural Language Processing, Transformer has been widely applied to tasks such as machine translation, summarization, question-answering, etc., achieving good performance in many domains.

Swin Transformer[10] is an image classification model based on the Transformer architecture, used for handling image tasks in computer vision. In comparison to traditional image classification methods like Convolutional Neural Networks (CNNs), Swin Transformer adopts a hierarchical attention mechanism that better models long-range dependencies in images. The core idea of Swin Transformer is the divide-and-conquer strategy, where the image is divided into a series of non-overlapping patches and a regular Transformer model is applied to these patches to process the entire image. This patch-based strategy helps reduce the computation and memory requirements of the model. Specifically, Swin Transformer consists of a series of hierarchical stages of encoders. Each encoder includes a set of patch-wise attention layers and a set of local window interaction layers. The patch-wise attention layers perform self-attention within each patch to capture spatial dependencies within the patch, while the local window interaction layers interact information between patches to model global context in the image.

Compared to traditional Transformers, Swin Transformer leverages the advantages of distributed computing by introducing patch-wise attention and local window interaction layers, thus better utilizing the computational resources and improving the computational and memory efficiency, making it applicable for handling large-scale images. Therefore, Swin Transformer exhibits competitiveness in image classification tasks and achieves leading performance on multiple benchmark datasets.

Advantages of Swin Transformer: Long-range dependency modeling: By adopting a hierarchical attention mechanism, Swin Transformer effectively models long-range dependencies in images, capturing global contextual information and improving the accuracy of image classification. Patch-wise strategy: Swin Transformer divides the image into non-overlapping patches and applies the Transformer model to each patch. This patch-wise strategy has lower computation and memory requirements, helping handle large-scale images and improving computational and memory efficiency. Scalability: With the patch-based approach, Swin Transformer can be horizontally expanded for processing images of different sizes and resolutions, demonstrating versatility and adaptability.

However, Swin Transformer also has certain limitations: Higher computational complexity: Although Swin Transformer reduces computation and memory requirements through the patch-based strategy, it still has higher computational complexity compared to traditional CNNs, requiring more computational resources and longer training time. Limited by local window interaction layers: The local window interaction layers in Swin Transformer are responsible for information exchange between patches, but this local interaction may not capture certain global contextual information, impacting the overall modeling capability of the model. Dependence on pre-training data: Similar to most deep learning models, Swin Transformer typically requires pre-training on large-scale datasets before achieving good performance, which may impose requirements on data collection and computational resources.

C. *DIOR*

As The DIOR dataset is a dataset used for object detection tasks in the computer vision field. It encompasses diverse images with variations in scale, viewpoint, and lighting conditions. DIOR stands for Deformable and Invariant Object Recognition, highlighting the robustness of object recognition against deformations and viewpoint changes. The dataset covers multiple object categories and scenes, such as humans, vehicles, buildings, and natural environments, and serves as an important benchmark for evaluating object detection algorithms.

However, the DIOR dataset also presents some challenges, including but not limited to the following: Class imbalance: The dataset exhibits an uneven distribution of samples across different categories, leading to potential performance issues in recognizing objects from minority classes. This can result in poorer detection performance for these minority object classes in real-world applications. Unclear object boundaries: In certain images of the DIOR dataset, object boundaries can be ambiguous due to factors like lighting variations or occlusions. Such challenges make object detection more difficult and may introduce errors at the boundaries of detected objects. Variations in viewpoint and scale: The DIOR dataset includes images with objects captured from different viewpoints and scales. Consequently, substantial changes in viewpoint or scale are observed in some cases. Algorithms designed for object detection need to possess robustness against object deformations and scale variations to mitigate adverse effects on detection accuracy.

It is crucial to consider these issues during algorithm design and training processes to overcome the limitations inherent in the DIOR dataset and enhance the robustness and performance of the models.

III. Dataset Analysis

A. *Imbalanced Sample Size Across Different Categories*

The DIOR[9] dataset exhibits class imbalance, which refers to uneven distribution of samples among different target categories. This means that certain target categories have significantly more instances in the dataset. Class imbalance can lead to performance bias in object detection algorithms, where the model tends to prefer majority classes and performs relatively poorly in accurately detecting minority class targets.

Imbalanced datasets can pose challenges during training, as the model may not have sufficient exposure to samples from minority classes, resulting in lower recognition performance on those classes. This can be problematic in real-world scenarios where accurate detection of all target classes is essential.

To mitigate the impact of class imbalance on model training, mosaic augmentation, which adds mosaic effects to images, can be employed. Mosaic augmentation is a data augmentation technique that involves adding mosaic tiles around the target to obscure parts of the image, creating new training samples. This technique helps balance the dataset and provide more training samples for minority classes.

B. *The Boundary Of The Target Is Relatively Blurry*

The problem of blurry object boundaries in the DIOR dataset refers to the insufficient clarity of the edge region of objects, making it difficult to accurately determine their precise boundaries, which can lead to poor localization and recognition performance, especially for complex objects. Accurately determining object boundaries is a critical step in object detection tasks, making it important to address this issue. The causes of this problem may be due to factors such as inappropriate selection of camera angle and distance or the relative position of the object with respect to the camera during object capture.

To address this issue, one can attempt to capture objects from multiple angles and distances to acquire clearer object boundaries and increase dataset diversity. When objects are partially obscured or occluded by other objects, restoration or filling algorithms can be used to improve object boundary clarity. Additionally, techniques such as mosaic or partial occlusion augmentation can generate more training samples, which can help models better understand and handle occlusion conditions. Uneven or intense lighting conditions can also cause blurry object boundaries, which can be improved by adjusting lighting conditions, using shadow reduction techniques, enhancing contrast and clarity of existing images, and using varied lighting conditions in training data to better adapt to different lighting situations.

C. *Object Images From Different Persoectives And Scales*

The problem of varying object viewpoints and scales in the DIOR dataset refers to the appearance of objects in different orientations and sizes in the images, caused by factors such as the position of the capturing device, the distance between the object and the camera, and the object's shape and pose. Accurately detecting and recognizing objects in object detection tasks requires handling and adapting to this diversity.

Including objects with multiple viewpoints and scales in the dataset can improve the model's generalization ability. By introducing images of objects with multiple viewpoints and scales in the dataset, the model can learn and adapt to variations in the object's appearance due to different orientations and sizes. This can be achieved by capturing objects from various angles and distances to obtain images of objects from different viewpoints. Additionally, including

object images with different scales in the training data can help the model adapt to objects of different sizes.

However, the existence of objects with multiple viewpoints and scales poses a challenge to the model's training.

IV. METHODS

A. YOLOv8_swin

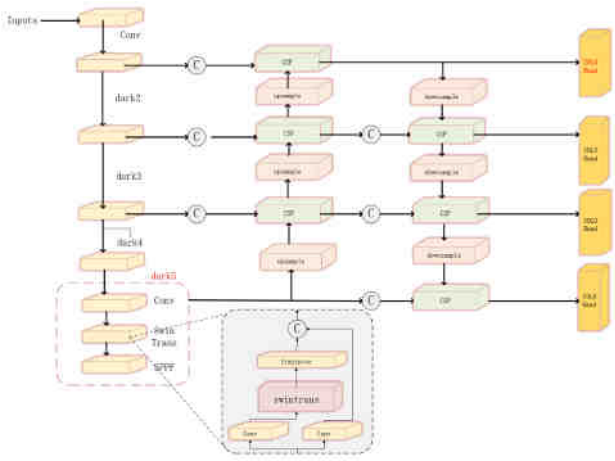
To address the issues in the DIOR dataset, this paper introduces an object detection algorithm called YOLO-SH based on YOLOv8. To achieve lightweight performance, modifications have been made to the YOLOv8 model by incorporating the Swin Transformer structure into the dark5 level of its Darknet backbone for improved feature extraction.

Swin Transformer is a deep learning model based on the Transformer architecture, which utilizes multi-stage attention mechanisms and cross-layer connections to effectively enhance the efficiency and accuracy of the model. Additionally, Swin Transformer offers better computational efficiency and smaller model size, making it suitable for practical applications. Considering the complex relationships between objects and their surrounding environments often present in remote sensing images, it is necessary to consider global information to perform better object detection. Therefore, the mechanism of multi-stage attention in the Swin Transformer structure can be used to weigh different levels and capture global information effectively.

B. YOLOv8_swin_head

In order to enhance the detection performance of small objects in the dataset, this paper incorporates YOLO heads based on YOLOv8, inspired by the TPH-YOLOv5[3] study. Three original YOLO heads from Darknet in YOLOv8 are extended to include an additional prediction head specifically designed for detecting tiny objects. While this introduces some additional computational burden, the overall model demonstrates better recognition performance for objects at various scales.

Figure 1. YOLO-SH



V. EXPERIMENT

A. Experiment Details

The experiments in this paper were conducted using Ubuntu 18.04.5 LTS as the operating system. The CPU used was an Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz, and the GPU model was an NVIDIA 4090. The Pytorch 1.8.0 framework was employed for training, with a CUDA version of 10.2 during the training process. The training images were of size 800*800 and in jpg format. In order to speed up the training process, a batch size of 8 was utilized and the learning rate was set to 0.01. Data augmentation was employed, as well as the use of mixed data augmentation. The cosine annealing algorithm was used to adjust the learning rate.

B. Evaluation Indicators

The experiments in this paper were conducted using the official DIOR dataset. The training and validation sets were classified in a 5:5 ratio within the official dataset. The model was trained following the dataset split provided by the official dataset and evaluated for Mean Average Precision (Map) using the official test set.

C. Experiment Result

The presented ablation experiments in this paper are shown in the corresponding table. After incorporating Swin Transformer structure, the model showed improvement in Map with respect to the baseline, and corresponding AP for each category also increased. YOLO head further increased the Map considerably and resulted in improved AP for each category. The performance of the proposed YOLO-SH model in comparison to other model approaches in the DIOR dataset is also shown in the table. On a faster FPS basis, the proposed model achieved a higher Map than the other network models, exhibiting significant enhancements in performance.

Table1. Comparison of model performance on DIOR dataset

Method	Speed(fps)	FLOPs(G)	mAP
CSFF[5]	9.9	\	68.0
FENet[6]	8.69	\	68.3
SSD(VGG) [7]	19.2	220.15	60.9
YOLOv3[4]	22.5	121.4	59.9
YOLOv4[8]	23.3	110.9	63.12
ours	29.02	28.8	77.02

Table2. Ablation experiment

Method	Madd(G)	FLOPs(G)	Memory	mAP
baseline	44.62	22.3	326MB	75.47
baseline + Swin	44.62	22.3	360.87MB	76.29
Baseline + Swin + haad	57.54	28.8	541MB	77.02

VI. CONCLUSION

With the increasing diversity of remote sensing applications, there is a growing number of remote sensing image datasets being generated. Against this background, remote sensing object detection has become a subject of significant interest. This paper focuses on analyzing the DIOR remote sensing dataset, particularly on three areas of interest: class imbalance in dataset samples, poor object boundary definition, and objects captured in different perspectives and scales. To address these issues, the paper proposes the YOLO-SH network, by incorporating the Swin Transformer structure and designing a specialized YOLO Head for small objects. Results of experimentation show that the proposed

model performs well on the DIOR dataset. The integration of the Swin Transformer structure into the YOLO-SH model brings powerful self-attention mechanisms and hierarchical feature representation capabilities. This enables the model to capture long-range dependencies and global contextual information, which is beneficial for various computer vision tasks such as object detection and image segmentation. The design of the YOLO Head specifically targets small objects, enhancing the model's accuracy in detecting them. Further research can explore ways to optimize this design for even better performance in small object detection. Applying the YOLO-SH model to different domains or types of remote sensing imagery would require transfer learning to adapt to different data distributions. Future investigations can focus on devising effective transfer learning strategies to improve the model's generalization ability and adaptability.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, You only look once: Unified, real-time object detection, in Proceedings of the IEEE conference on computer vision and pattern recognition (2016) pp. 779–788.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017)

[3] X. Zhu, S. Lyu, X. Wang and Q. Zhao, Tph-yolov5: Improved yolov5 based on trans former prediction head for object detection on drone-captured scenarios, in Proceed ings of the IEEE/CVF international conference on computer vision (2021) pp. 2778– 2788.

[4] J. Redmon and A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767 (2018).

[5] G. Cheng, Y. Si, H. Hong, et al., “Cross-scale feature fusion for object detection in optical re mote sensing images,” IEEE Geoscience and Remote Sensing Letters 18(3), 431–435 (2020). [doi:10.1109/LGRS.2020.2975541].

[6] G. Cheng, C. Lang, M. Wu, et al., “Feature enhancement network for object de tection in optical remote sensing images,” Journal of Remote Sensing 2021 (2021). [doi:10.34133/2021/9805389].

[7] W. Liu, D. Anguelov, D. Erhan, et al., “Ssd: Single shot multibox detector,” in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11– 14, 2016, Proceedings, Part I 14, 21–37, Springer (2016). [doi:10.1007/978-3-319-46448- 0 2].

[8] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “Yolov4: Optimal speed and accuracy of ob ject detection,” arXiv preprint arXiv:2004.10934 (2020). [doi:10.48550/arXiv.2004.10934].

[9] K. Li, G. Wan, G. Cheng, et al., “Object detection in optical remote sensing images: A survey and a new benchmark,” ISPRS journal of photogrammetry and remote sensing 159, 296–307 (2020). [doi:10.1016/j.isprsjprs.2019.11.023].

[10] Liu Z , Lin Y , Cao Y ,et al.Swin Transformer: Hierarchical Vision Transformer using Shifted Windows[J]. 2021.DOI:10.48550/arXiv.2103.14030.