TSM-YOLO: Multi-expert Mechanism for Small Object Detection in UAV Perspectives

Ziming Zhang

***Abstract*—Currently, deep learning-based object detection methods have shown remarkable performance on traditional datasets. Nevertheless, when encountering small objects captured from an Unmanned Aerial Vehicle (UAV) perspective, challenges arise in the form of reduced accuracy, elevated false detection rates, and missed detections. To address these issues, we introduce a novel dual-branch backbone object detection algorithm that synergizes the strengths of both Transformer and Convolutional Neural Network (CNN).In the feature extraction stage , our algorithm addresses the challenges posed by the uneven distribution of small objects and imbalanced useful information for detection tasks. Specifically, the primary branch employs a CNN Backbone to capture multi-scale features and the secondary branch effectively captures high-resolution features enriched with global information. Furthermore, we introduce a Cross-attention Module (CAM) feature fusion module, which aids PANet in seamlessly blending high- and low-frequency information. For the classification and detection tasks, we adopt YOLOV8's decoupled Head and post-processing methodologies, ensuring compatibility and efficiency. Meanwhile, we propose a Co-Upcycling training strategy to optimize the training of our multi-expert modules. The efficacy of our proposed method is rigorously evaluated on the Visdrone-2019 dataset. The experimental outcomes reveal that our approach surpasses the YOLOv8-s benchmark, achieving improvements of 4.2% and 3.1% in mAP50 and mAP50-95, respectively.**

***Index Terms*— MoE, SOD, Transformer, Visdrone.**

I. INTRODUCTION

Currently, deep learning-based object detection has emerged as a pivotal research area, significantly advancing both the precision and efficiency of object detection tasks. Nonetheless, the detection of small objects remains a formidable challenge. Small objects, inherently characterized by their diminutive sizes and limited representation within images, often suffer from a dearth of available and discriminative features. Additionally, current state-of-the-art detectors frequently exhibit suboptimal generalization performance when confronted with these diminutive targets [1].This predicament stems primarily from the fact that small objects are dispersed across diverse regions of the image, where heterogeneous backgrounds and occlusions contribute to blurred and indistinct appearances. Consequently, these small objects are highly susceptible to the influence of vast background regions and image noise. Moreover, the scarcity of datasets tailored specifically for training small object detection in complex scenarios exacerbates this issue. Nevertheless, given the wide range of applications for small object detection, including autonomous driving, traffic surveillance, defect detection, and aerial image analysis, it has become one of the most active and pressing research topics in the field of object detection tasks.

This paper presents a detection algorithm specifically designed for small objects, based on the YOLO paradigm. The primary branch of the Backbone continues to employ the YOLOV8 Backbone, while the auxiliary branch incorporates the Aggregated Attention mechanism from TransNeXt [6], alongside the GLU Channel Mixer, to propose a TSMBlock expert module that adaptively extracts crucial information of objects within the image, neglecting non-target regions. Furthermore, we introduce a CAM module grounded on the Cross Attention architecture, which assists PANet in performing multi-scale feature fusion. This fusion process aims to retain details and texture information that may be lost during downsampling in the Backbone stage. For classification and detection tasks, we maintain the decoupled Head and Anchor Free mechanism from YOLOV8. Finally, we evaluate our approach using the VisDrone-2019 dataset.

II. RELATED WORK

Research on small target detection is generally approached from two angles: multi-scale feature representation and contextual information [1].

A. Multi-scale feature representation

The depth of image features in deep learning determines the amount of information they contain. Shallow features with higher resolution are rich in detailed and textural information, while deep features with lower resolution exhibit stronger semantic and task-specific information. [2] employs multi-scale fusion by downsampling shallow features and upsampling deep features, stacking them in the channel dimension to obtain a single-resolution feature map that combines strong semantic with abundant location information, and predictions are made at this resolution. PAN[3] introduce multi-scale feature fusion and prediction, accommodating different object sizes and significantly enhancing the accuracy of object detection. [4] implementing a dual-path network to retain high-resolution features conducive to small object detection as much as possible. [5] designs a global context network and a pyramid local context network to extract global and local information, respectively, along with a spatial and scale-aware attention module that guides the network to focus on more informative regions and appropriate image feature scales. Additionally, [6] utilize higher-resolution feature maps to extend additional detection heads specifically for small objects.

B. Contextual information

Given the challenging nature of extracting features from

Manuscript received March 14, 2025
Ziming Zhang, School of computer and technology,Tiangong University, Tianjin,China

small target, context information modeling capability with adaptive perception are particularly crucial. The progression from C3, E-ELAN to the C2F architecture of YOLO[7] showcases the adoption of efficient connection patterns and the introduction of more branches, which not only enhance computational performance but also skew the feature extraction process towards capturing complex information in images from multiple perspectives. However, in the context of convolutions, their inductive biases primarily manifest as locality, translation invariance, and weight sharing [8]. These attributes are not well-suited to the heterogeneous distribution of objects and backgrounds, as well as the uneven distribution of salient information in small object datasets. SE[9] argues that the weights of different channels should be adaptively allocated, proposing a method to selectively emphasize informative features and suppress unnecessary ones by learning from global information. Furthermore, the introduction of ViT (Vision Transformer) has enabled global modeling of images. To address the quadratic complexity issue inherent in the self-attention mechanism, PVT[10] proposes SRA, which aims to reduce the feature space dimension of the entire model by spatially downsampling the keys and values in the attention mechanism. Additionally, local attention mechanisms have been proposed by Swin-T[11], CSwin-T[12], and others, offering alternative approaches to tackle the challenges posed by small objects.

III. METHODS

A. TSM-YOLO

In our approach, we introduce an auxiliary branch in the backbone section, improved TSMNet based on TransNeXt. The Aggregating Attention mechanism[13] of TransNeXt is retained in the Block stage to capture both fine-grained window-level information and coarse-grained global information, as illustrated in Fig.3. For the feedforward layer, in view of the diversity in object types and shapes encountered in small object detection tasks, we adopt a multi-expert paradigm. We propose utilizing gating units to adaptively select the appropriate feedforward experts for each image patch, dynamically modulating the extraction intensity and processing approach for different objects and background information. In the Neck stage, we first employ CAM to extract high-frequency texture information with global context from the auxiliary branch, subsequently fusing features across different scales through PANet. Ultimately, the feature maps are forwarded to the Decoupled Head for detection tasks.The TSM-YOLO architecture is shown in Fig.1.

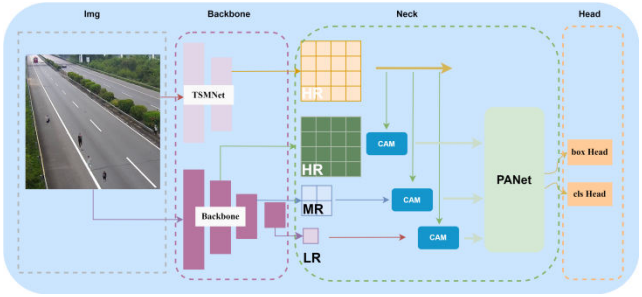

Fig.1. TSM-YOLO model architecture

B. TSMNet Auxiliary Branch

Biological vision is characterized by heightened sensitivity to features near the current visual focus, with lesser concern for distant features. Fig.2 shows the visual characteristics simulated by different attention mechanisms. Moreover, this property of biological vision remains consistent across pixels at any position within an image, implying pixel-level translational equivalence. In contrast, local attention based on window partitioning does not treat pixels at the window edges and center as equivalent in terms of attention, revealing a notable discrepancy.

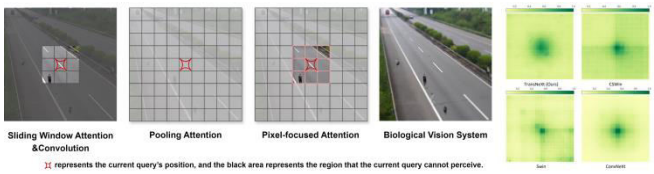

Fig.2. Different attention mechanisms show areas of concern

TransNext proposed a focused attention mechanism, enabling fine-grained perception in the vicinity of each Query while maintaining a coarse-grained awareness of global information. To achieve the inherent pixel-level translational equivalence in eye movements, a dual-path design is adopted, comprising a Query-centric sliding window attention and a pooling attention. As shown in Fig.3. The blue area is window Attention, the orange area is global pooling Attention, and the yellow area is the fusion of the two types of attention.

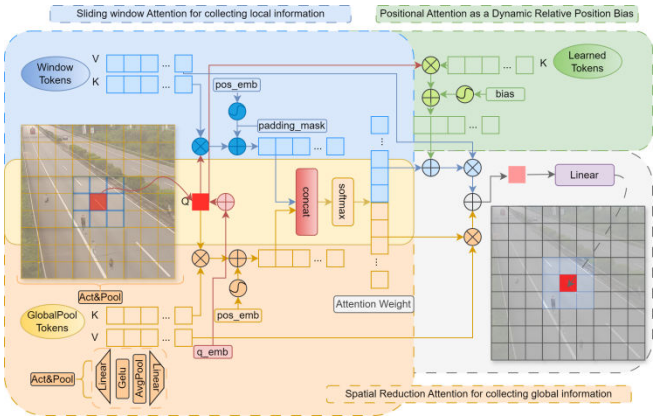

Fig.3. Aggregated Attention framework.

Addressing the challenges posed by small object detection tasks, such as uneven distribution of small objects within images, excessive background clutter with redundant information, and the inherent multi-morphological and multi-scale nature of the targets, we devise a feature extraction module, TSMBlock, equipped with a self-adaptive expert-selective channel mixer[14]. As shown in Fig.4. This module harnesses a Router unit to adaptively select the appropriate GLU expert for each image patch. Considering consistency of tasks, we adopt a universal expert, the Shared GLU, to process all patches, and subsequently weight its results with those from the remaining experts. The modular design is illustrated in the accompanying figure.

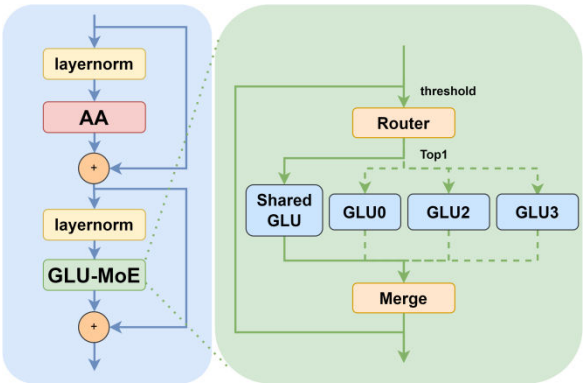

Fig.4. TSMBlock

The TSMNet presented in this paper remains grounded in the TransNeXt paradigm, with the original blocks substituted by TSMBlocks. We have opted for the tiny version, and amend to the configuration. Specifically, to align with YOLOV8s, the value of C is set to 64, and Ni is initially defined as [2,15,15,2],as illustrated in the accompanying Fig.5. However, in this paper, we solely utilize the first two stages, configuring Ni as [2,10], with mlp ratio set to [4,4], window size specified as [3,3], and head dimensions adjusted to [2,4].

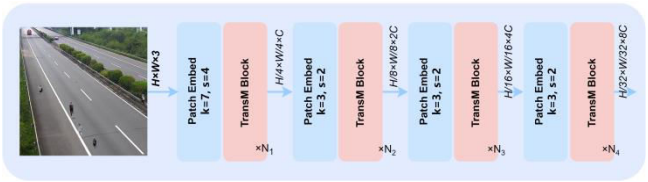

Fig.5. TSMNet

C. CAM

After downsampling, as the image size decreases, the localization information crucial for detecting small objects is also lost. Consequently, the detection of small objects places greater emphasis on high-resolution feature maps that possess global information. To mitigate the potential impacts arising from the integration of modules with distinct characteristics, we propose a CAM approach. According to Fig.6. This method leverages the Cross Attention mechanism, where the multi-scale features (LR) from the main branch serve as Queries (Q), while the high-resolution features (HR) from the auxiliary branch function as Keys (K) and Values (V). This allows the multi-scale features extracted by the convolutional backbone to acquire additional global and detailed information from the high-frequency features extracted by the corresponding auxiliary backbone, facilitating subsequent processing by PANet. As shown in formula (1), it encapsulates the process by which low frequency features extract information from high frequency features. Each Query (Q) is confined to the corresponding sub-region of the high-resolution features (HR), thereby maintaining efficiency. This design strikes a balance between preserving the richness of low-frequency feature details and computational feasibility.The calculation formula of CAM is (1).

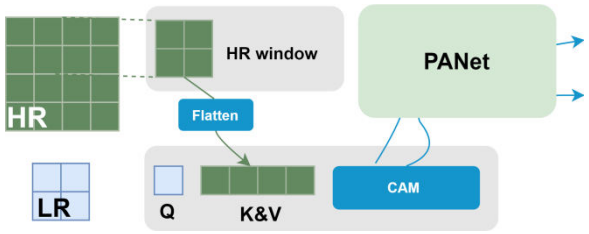

Fig.6. CAM with PANet

$$\mathrm{CAM}(\mathrm{LR},\mathrm{HR}) = \mathrm{Q_{LR}} + \mathrm{softmax}(\mathrm{Q_{LR}}@\mathrm{K_{HR}})@\mathrm{V_{HR}} \quad (1)$$

D. Co-upcycling MoE

During the training, we encountered the challenge of model convergence difficulties, particularly when training the newly introduced Mixture-of-Experts (MoE) blocks from scratch. To address this issue, we adopted the Co-Upcycling approach, which incorporates a two-stage training paradigm. According to Fig.7. In the first stage, we trained the TSMBlock without the GLU-MoE integration. Subsequently, in the second stage, we initialized the GLU module incorporating the MoE blocks using the pre-trained GLU from the first stage. This strategy facilitated the convergence and performance of the model by leveraging the learned representations from the initial training phase.

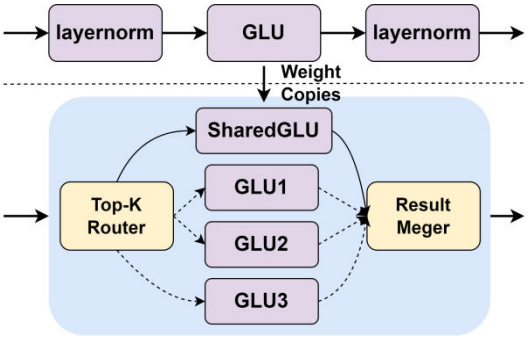

Fig.7. Co-Upcycling Training Strategies

IV. EXPERIMENT

A. Experiment description

To evaluate the performance of the proposed TSM-YOLO model for small object detection from an unmanned aerial vehicle (UAV) perspective, we utilized the Visdrone2019 dataset. Firstly, we conducted a comparative analysis of the latest detection algorithms and ours, focusing on metrics such as detection accuracy, speed, computational complexity, and parameter count. Secondly, ablation studies were performed to test the TSM-YOLO model with various modifications, examining the impact of different enhancement strategies on its performance.

During training, the input images are resized to 640 × 640 pixels, with the training process spanning 400 epochs, each batch containing 16 images. Data augmentation techniques are employed to enhance the diversity of target samples. All experiments are conducted using PyTorch on a GeForce GTX 3090 GPU.

B. *VisDrone2019*

The VisDrone2019 dataset stands out as a renowned benchmark for object detection from a UAV perspective, widely employed in UAV-based object detection tasks. It embodies both diversity and richness, encompassing 288 video clips, composed of 261,908 frames and 10,209 static images. These images and videos span a broad spectrum of scenes, covering different locations, environments, objects, and densities. The VisDrone2019 dataset boasts extensive annotation information, including over 2.6 million bounding boxes, encompassing a wide range of common objects. Fig.8 illustrates the VisDrone2019 dataset, shown on the left during the day and on the right at night.

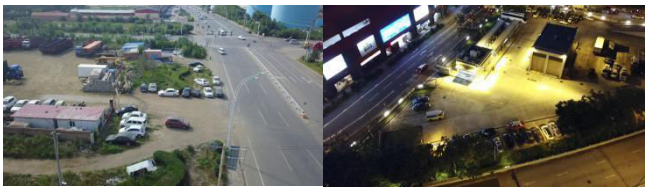

Fig.8. A sample of the Visdrone2019 dataset

C. *Experimental result*

Table.1 presents a comparative analysis of the state-of-the-art small object detection algorithms against the TSM-YOLO in terms of various evaluation metrics. The accuracy is evaluated using Recall, mAP50, and mAP50-95, expressed as percentages.

It is observed that TSM-YOLO exhibits only a marginal increase in FLOPs compared to YOLOv8s, yet its mAP surpasses YOLOv8s by 4 points. This enhancement can be attributed to the multi-expert mechanism employed in the TSM Block, which, despite housing four GLU units, activates only two of them, enabling an elegant utilization of its extensive weight parameters.

Table 1. Comparative experimental results

Method	mAP50/%	mAP50-95/%	Recall/%	Params/M	FLOPs/G
YOLOv7-t	37.8	22.7	38.1	6.01	13.1
YOLOv8-s	40.3	24.2	40.6	11.2	28.6
YOLOv9-s	40.1	25.1	40.8	7.2	26.7
RT-DETR-R18	35.3	23.8	35.4	20.0	60.0
TSM-YOLO	**44.3**	**27.1**	**43.1**	**23.9**	**37.8**

To gain a more intuitive understanding of the impact of various enhancement techniques on model performance, we conducted an ablation study. Specifically, while maintaining the architecture of YOLOv8s unchanged, we incrementally incorporated the TransNeXt auxiliary backbone, TSMNet auxiliary backbone, the CAM auxiliary module, and the Co-Upcycling Mixture of Experts (MoE) training strategy into the model's training process. The results are shown in table 2.

Table 2. Ablation experiment results

Method	mAP50/%	mAP50-95/%	Recall/%
Base	40.3	24.2	40.6
TransNext+CAM	41.3	25.2	41.1
TSMNet+CAM	43.9	25.9	42.5
All	44.3	27.1	43.1

V. CONCLUSION

To improve the accuracy of small object detection, this paper introduces an approach that integrates an Adaptive Selection Feature Module, termed TSMBlock, alongside an auxiliary feature fusion mechanism utilizing the Context Attention Module (CAM). This methodology selectively extracts features of objects that require heightened attention, adhering to the principles of biological visual attention mechanisms. Simultaneously, it accommodates morphological variations among multiple targets by adaptively selecting distinct feed forward experts to blend channel information. Experimental results indicate that our method achieves superior accuracy in small object detection tasks from an UAV perspective.

REFERENCES

[1] Chen G, Wang H, Chen K, et al. A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal[J]. IEEE Transactions on systems, man, and cybernetics: systems, 2020, 52(2): 936-953.

[2] Li K, Zhang W, Yu D, et al. HyperNet: A deep network for hyperspectral, multispectral, and panchromatic image fusion[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 188: 30-44.

[3] Shi D. Transnext: Robust foveal visual perception for vision transformers[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 17773-17783.

[4] Zhou Q, Shi H, Xiang W, et al. DPNet: Dual-path network for real-time object detection with lightweight attention[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024.

[5] Yang Y, Guo M Q, Zhu Q. CADNet: Top-down contextual saliency detection network for high spatial resolution remote sensing image shadow detection[C]//2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. IEEE, 2021: 4075-4078.

[6] Wang G, Chen Y, An P, et al. UAV-YOLOv8: A small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios[J]. Sensors, 2023, 23(16): 7190.

[7] Hussain M. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection[J]. Machines, 2023, 11(7): 677.

[8] Rekavandi A M, Rashidi S, Boussaid F, et al. Transformers in small object detection: A benchmark and survey of state-of-the-art[J]. arXiv preprint arXiv:2309.04902, 2023.

[9] Wu T, Dong Y. YOLO-SE: Improved YOLOv8 for remote sensing object detection and recognition[J]. Applied Sciences, 2023, 13(24): 12977.

[10] Wang W, Xie E, Li X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 568-578.

[11] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[12] Dong X, Bao J, Chen D, et al. Cswin transformer: A general vision transformer backbone with cross-shaped windows[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 12124-12134.

[13] Shi D. Transnext: Robust foveal visual perception for vision transformers[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 17773-17783.

[14] Jiang A Q, Sablayrolles A, Roux A, et al. Mixtral of experts[J]. arXiv preprint arXiv:2401.04088, 2024.