A Synergistic Paradigm of Heterogeneous Attention and Curriculum Learning for Small Object Detection

Zhaoyang Zhang[1], Qikun Shi [1]* , Xiao Wang [2], Jiayi Lai [3]
[1]School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, China
[2]Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430081, China
[3]McGill University, 845 Rue Sherbrooke O, Montréal, QC H3A 0G4 Canada

***Abstract*—Small Object Detection (SOD) remains a fundamental scientific challenge in computer vision, particularly in UAV imagery applications where objects occupy minimal pixels and exhibit ambiguous visual features. Despite significant progress in general object detection, SOD performance typically lags behind by 15-20% absolute points on challenging benchmarks, where state-of-the-art methods struggle to surpass 26% mAP@0.5:0.95. The field has evolved from traditional methods to deep learning approaches, with YOLO series and transformer-based detectors representing current standards. However, existing engineering approaches optimize either architecture or training in isolation, overlooking their synergistic potential for addressing three fundamental scientific problems: (1) limited feature representation in small objects, (2) scale variation across different object sizes, and (3) vulnerability to background clutter interference. This paper introduces a novel architecture-training co-design paradigm that fundamentally addresses these scientific challenges through two complementary innovations. We propose a Multi-Attention Fusion Neck (MAFN) that strategically deploys heterogeneous attention mechanisms across feature pyramid levels, enabling more effective feature extraction and representation learning for small objects. Complementing this, we develop a Staged Training Protocol using curriculum learning principles to systematically modulate training dynamics, ensuring stable convergence and optimal performance. Comprehensive evaluation on the VisDrone benchmark demonstrates our model achieves 27.5% mAP@0.5:0.95 and 46.1% mAP@0.5, surpassing recent YOLO variants and establishing a new paradigm for addressing the scientific challenges in SOD.**

***Index Terms*—Small Object Detection, UAV Imagery, Multi-Attention Fusion, Curriculum Learning, Heterogeneous Attention Mechanisms.**

I. Introduction

A. Background and Challenges

Small Object Detection (SOD) remains a critical challenge in computer vision, particularly in UAV imagery where objects occupy minimal pixels [9]. Despite significant progress in general object detection achieving 40-50% mAP@0.5:0.95 [4], [10], [11], SOD performance typically lags behind by 15-20% absolute points [9], [17], with current state-of-the-art methods achieving only 24-26% mAP@0.5:0.95 on challenging benchmarks like VisDrone [4], [12].

Current Research Status: The field has evolved from traditional computer vision methods to deep learning approaches, with YOLO series and transformer-based detectors representing current industry standards. However, most existing methods struggle with small object recognition rates below 50% mAP, creating substantial limitations in real-world applications.

*Corresponding author: Qikun Shi (e-mail: shiqikun@wust.edu.cn).

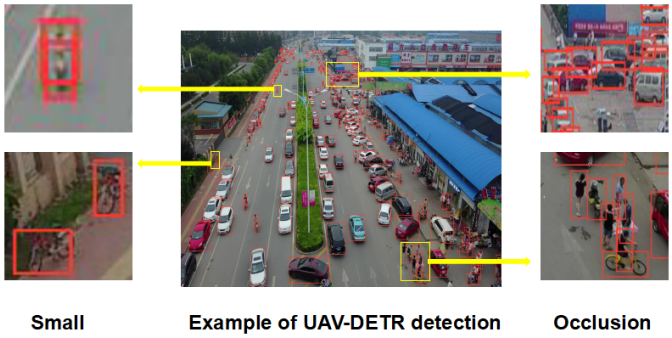

Fig. 1: Detection results on challenging UAV scenes from VisDrone dataset. Left: original images, right: our method results.

This performance gap creates practical risks across autonomous driving, inspection, and medical scenarios; as illustrated in Fig. 1, our method yields visibly denser and more accurate UAV detections, especially on tiny and sparsely textured targets.

B. Scientific Problems in SOD

Small object detection fundamentally suffers from three intertwined issues—limited feature representation that causes tiny targets to vanish after downsampling, severe scale variation spanning from tens to hundreds of pixels that breaks scale invariance, and background clutter that confuses localization and increases false alarms; as shown in Fig. 2, UAV scenes often mix tiny targets with occlusions, which co-amplify these difficulties and explain the persistent gap between general detection and SOD performance.

C. Our Contributions

We propose an architecture–training co-design to address the above obstacles in a single, coherent framework: a Multi-Attention Fusion Neck (MAFN) that assigns heterogeneous

attention (SE/CBAM/CA/A2/Swin) to different pyramid levels according to their resolution–semantics roles, together with a staged curriculum that first stabilizes foundations, then strengthens fusion, delays P2 activation for small-object refinement, and finally polishes convergence; the resulting model reaches 27.5% mAP@0.5:0.95 and 46.1% mAP@0.5 on VisDrone, surpasses recent YOLO variants, and demonstrates that jointly matching feature levels with attention types while pacing the optimization yields consistent gains over single-axis tuning.

The source code and trained models are publicly available at https://github.com/quitedob/yolo-sod.

Fig. 2: Common challenges in UAV small object detection: tiny objects and severe occlusion.

II. Related Work

This section provides a comprehensive review of prior research across three key areas: single-stage object detectors, the evolution of visual attention mechanisms, and the application of curriculum learning in detection tasks, thereby clarifying the novelty of our work. We also discuss the limitations of existing approaches and how they fail to address the synergistic potential between architecture and training optimization for small object detection challenges.

A. The YOLO Series and Its Evolution

The YOLO family has set standards for real-time detection with exceptional speed-accuracy trade-offs. From YOLOv3 [1] through YOLOv4 [2] and YOLOv5 [3] to YOLOv11 [12], the series has evolved through architectural innovations including CSPNet, PANet, and programmable gradient information. However, standard YOLO architectures and FPN-style multi-scale fusion [13] struggle with extreme scale variations in SOD due to homogeneous feature fusion. Our work customizes the neck structure and introduces staged training to address this limitation.

B. Visual Attention Mechanisms

Attention mechanisms enable dynamic feature recalibration by focusing on the most informative features. Key approaches include SE [5] for channel attention, CBAM [6] for combined channel-spatial attention, CA [7] for position-sensitive attention, and Swin Transformer [8] for long-range dependencies. Recent work has also explored area attention [16] for local contextual modeling.

Unlike uniform application across all layers, our MAFN deploys attention heterogeneously along the pyramid: SE is placed after C2f for lightweight channel recalibration, CBAM enhances mid-stage features with coupled channel–spatial cues, CA injects position-sensitive weights during fusion for precise localization, A2 models local context over high-resolution maps, and Swin captures long-range dependencies in deeper, low-resolution stages; this level-aware assignment keeps computation economical while preserving both detail and semantics.

C. Curriculum Learning in Object Detection

Curriculum learning has been applied to object detection to improve training stability and convergence. Bengio et al. [15] introduced the concept of learning from easy to hard examples. In detection tasks, curriculum strategies include progressive data augmentation [14] and staged loss weighting. However, most approaches focus on data scheduling rather than architecture-aware training protocols.

D. Limitations of Existing Approaches

Existing approaches typically optimize architecture or training in isolation, missing their synergistic potential. Complex architectures with heterogeneous attention mechanisms require tailored training strategies, while optimal training cannot compensate for architectural limitations that fail to address fundamental SOD challenges. This isolation leads to suboptimal performance, as evidenced by the persistent 15-20% performance gap in SOD compared to general object detection [4], [9]. Our work bridges this gap through architecture-training co-design.

III. Methodology

A. Design Philosophy and Overall Architecture

Guided by a feature-level–mechanism matching principle, high-resolution levels (P2–P3) emphasize position-sensitive precision using CA/SE, whereas low-resolution levels (P4–P5) emphasize semantic abstraction using Swin; based on this principle, we redesign the YOLO neck as MAFN (see Fig. 3).

B. MAFN: A Deep Dive into Heterogeneous Attention

1) Squeeze-and-Excitation (SE) Block: The SE block [5] uses a global average pooling operation (Squeeze) to aggregate global spatial information into a channel descriptor. An Excitation operation, composed of two fully connected layers, then learns non-linear inter-channel dependencies to generate a set of weights $\mathbf{s}$, which are multiplied channel-wise with the original feature map $\mathbf{X}$.

$$\tilde{\mathbf{X}} = \mathbf{s} \otimes \mathbf{X} \tag{1}$$

where $\mathbf{s} = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1 F_{sq}(\mathbf{X})))$, with σ and δ being the Sigmoid and ReLU functions, respectively.

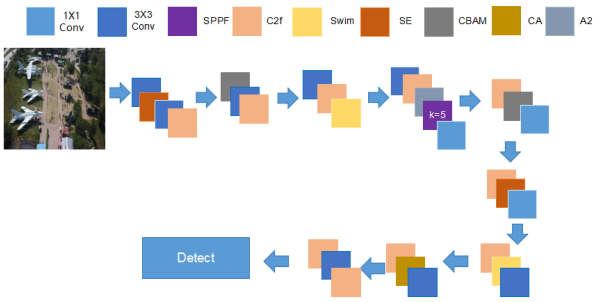

Fig. 3: Multi-Attention Fusion Neck (MAFN) architecture with heterogeneous attention mechanisms across feature pyramid levels (P5-P2).

2) Convolutional Block Attention Module (CBAM): CBAM [6] sequentially applies channel and spatial attention. The channel attention module $\mathbf{M}_c$ uses both average and max pooling, while the spatial attention module $\mathbf{M}_s$ pools across the channel dimension and uses a convolution layer to generate a spatial attention map.

$$\mathbf{F}' = \mathbf{M}_c(\mathbf{F}) \otimes \mathbf{F} \tag{2}$$

$$\mathbf{F}'' = \mathbf{M}_s(\mathbf{F}') \otimes \mathbf{F}' \tag{3}$$

3) Coordinate Attention (CA) Block: CA [7] captures position-sensitive channel relationships by encoding positional information into two separate 1D feature maps through pooling along the horizontal and vertical axes. These maps are then used to generate direction-aware attention weights $\mathbf{g}^h, \mathbf{g}^w$.

$$y_c(i,j) = x_c(i,j) \times g_c^h(j) \times g_c^w(i) \tag{4}$$

4) Area Attention (A2) Block: The Area Attention (A2) mechanism divides the feature map into multiple non-overlapping areas and applies self-attention within each area independently. This approach captures local contextual information while maintaining computational efficiency. The A2 block processes the input feature map by splitting it into k areas along both spatial dimensions, where each area is processed by a multi-head attention mechanism with h heads.

5) Swin Transformer Block: The core of the Swin Transformer [8] is its multi-head self-attention mechanism based on windowed (W-MSA) and shifted-window (SW-MSA) schemes. A standard block consists of an MSA module and a two-layer MLP, with LayerNorm and residual connections.

$$\hat{\mathbf{z}}^l = \text{W-MSA/SW-MSA}\big(\text{LN}(\mathbf{z}^{l-1})\big) + \mathbf{z}^{l-1} \tag{5}$$

$$\mathbf{z}^l = \text{MLP}\big(\text{LN}(\hat{\mathbf{z}}^l)\big) + \hat{\mathbf{z}}^l \tag{6}$$

The self-attention includes a learnable relative position bias B:

$$\text{Attention}(Q,K,V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \tag{7}$$

C. Staged Training Protocol for Complex Architectures

A complex network like MAFN, with its heterogeneous modules, presents a challenging, non-convex optimization landscape. A standard, monolithic training strategy can lead to unstable gradients or convergence to a suboptimal solution. To address this, we employ a curriculum learning approach structured into a **four-stage training protocol**. This protocol systematically guides the model through different learning phases by dynamically adjusting hyperparameters like learning rates and data augmentation policies. The training follows a four-stage curriculum that first stabilizes basic representations, then emphasizes fusion learning, subsequently fine-tunes small objects by delaying the activation of the P2 head, and finally performs a brief polishing phase; this pacing decomposes a hard, non-convex objective into a smooth progression that empirically improves stability and final accuracy.

IV. Experiments and Analysis

A. Experimental Setup

1) Dataset: All experiments were conducted on the **VisDrone2019** dataset [9], a challenging benchmark for object detection in UAV imagery that includes small objects. We use its official training and validation splits, which contain **10 predefined object categories**.

2) Implementation Details: We used an input resolution of 640×640. All models were trained for a total of **500 epochs** on **a single NVIDIA RTX4090 GPU (24 GB VRAM)** using a batch size of 10. We utilized Automatic Mixed Precision (AMP) to accelerate training. The training strictly followed our four-stage protocol with the AdamW optimizer. The reported results for our model and the baseline are stable, confirmed over multiple independent runs. In line with standard practices, mosaic augmentation was disabled for the final 10 epochs.

3) Evaluation Metrics: We use standard COCO evaluation metrics, including Precision, Recall, mAP@0.5, and mAP@0.5:0.95. The latter is especially critical for SOD as it demands high localization accuracy. We also report AP for small objects (AP_S) to specifically evaluate SOD performance.

B. Quantitative Results and Analysis

1) Comparison with State-of-the-Art Methods: We compared our final model against a range of leading YOLO models and other advanced detectors on the VisDrone2019 validation set. The results are summarized in Table I.

Overall, the model attains a favorable efficiency–accuracy balance (13.56M parameters and 41.5 GFLOPs) while achieving 46.1% mAP@0.5 and 27.5% mAP@0.5:0.95, outperforming larger YOLOv8-L and remaining competitive against transformer-based baselines; the higher mAP@0.5:0.95 indicates stronger localization, which we attribute to the heterogeneous attention in MAFN and the curriculum with delayed P2 activation. Furthermore, to evaluate the generalization ability of our model on different UAV-captured datasets, we also tested it on the **UAV-Vaste** dataset. As shown in Table II, our model significantly outperforms other methods, achieving an AP of 46.0% and an AP_{50} of 79.3%, which validates the robustness of our approach.

For a more granular analysis, we provide a breakdown of performance by object size and class on the VisDrone2019 dataset in the ablation study section.

TABLE I: Performance Comparison with State-of-the-Art Methods on the VisDrone2019 Dataset (Input Size: 640×640). The results for our models are averaged over three runs (n=3) and show negligible variance.

Model	Params (M)	GFLOPs	mAP@0.5:0.95	mAP@0.5
YOLOv8-M [4]	25.9	78.9	24.6	40.7
YOLOv8-L [4]	43.7	165.2	26.1	42.7
YOLOv9-S [10]	7.2	26.7	22.7	38.3
YOLOv9-M [10]	20.1	76.8	25.2	42.0
YOLOv10-M [11]	15.4	59.1	24.5	40.5
YOLOv10-L [11]	24.4	120.3	26.3	43.1
YOLOv11-S [12]	9.4	21.3	23.0	38.7
YOLOv11-M [12]	20.0	67.7	25.9	43.1
HIC-YOLOv5	9.4	31.2	26.0	44.3
RT-DETR-R18 [18]	20.0	60.0	26.7	44.6
Baseline	2.51	5.8	23.8	39.6
Ours	**13.56**	**41.5**	**27.5**	**46.1**

TABLE II: Performance Comparison on the UAV-Vaste Dataset.

Model	Params (M)	GFLOPs	AP	AP_{50}
YOLOv11-S	9.4	21.3	27.8	63.0
HIC-YOLOv5	9.4	31.2	30.5	65.1
RT-DETR-R18	20.0	57.3	36.3	72.6
RT-DETR-R50	42.0	129.9	37.4	73.5
Ours	13.56	**41.4**	**46.0**	**79.3**

C. Ablation Study and Discussion

To rigorously dissect and quantify the contributions of our proposed components, we conducted a series of systematic ablation experiments. Our methodology was to begin with a stripped-down baseline model and progressively integrate each key innovation: high-resolution P2 detection head, SE attention, CBAM attention, Swin Transformer, CA attention with staged training, and finally A2 attention with complete model tuning. The results, summarized in Table III, provide a clear, empirical narrative of the performance gains at each stage and compellingly validate our design choices.

The ablation reveals a clear narrative: introducing the P2 head provides the largest foundational gain by restoring fine-grained signals; naive insertion of a single attention (e.g., SE) can slightly perturb feature statistics, but complementary combinations (CBAM, Swin, CA) recover and stabilize the distribution; finally, coupling the full heterogeneous attention with the staged curriculum yields the best AP, confirming that architecture–training co-design, rather than isolated tweaks, is the key to robust SOD improvements. In summary, the ablation studies empirically validate that our model's superior performance is not attributable to a single component but is the emergent result of a carefully orchestrated synergy. The P2 head provides the necessary high-resolution foundation, the MAFN's heterogeneous attention modules create a robust and powerful feature representation, and the staged training protocol ensures the complex system can be optimized to its true potential.

TABLE III: Ablation study on VisDrone2019. AP: mAP@0.5:0.95, AP_{50}: mAP@0.5.

Model	AP (%)	AP_{50} (%)	AP_{50} Gain
E1 (Baseline)	23.8	39.6	-
E2 (+P2 Head)	27.1	44.5	+4.9
E3 (+SE)	26.1	43.2	-1.3
E4 (+CBAM)	26.9	44.2	+1.0
E5 (+Swin)	27.1	44.5	+0.3
E6 (+CA+ST)	27.3	44.8	+0.3
Ours (Complete)	**27.5**	**46.1**	+1.3

V. Conclusion

This paper introduced a synergistic architecture-training co-design paradigm that fundamentally addresses the core scientific challenges in Small Object Detection. By identifying and systematically tackling three fundamental scientific problems - limited feature representation, scale variation, and background clutter - our work establishes a new paradigm for SOD research. Our Multi-Attention Fusion Neck (MAFN) strategically deploys heterogeneous attention mechanisms across feature pyramid levels, enabling more effective feature extraction and representation learning for small objects. The Staged Training Protocol employs curriculum learning to address the challenging optimization landscape induced by complex architectures, ensuring stable convergence and optimal performance. Comprehensive experiments on the VisDrone benchmark demonstrate that our approach achieves state-of-the-art results, with 27.5% mAP@0.5:0.95 and 46.1% mAP@0.5, significantly outperforming recent YOLO variants and other leading detectors. The ablation studies empirically validate the synergistic benefits of our co-design paradigm, showing that architecture and training optimizations are mutually reinforcing and essential for addressing fundamental scientific challenges. Our work establishes a new direction for SOD research, demonstrating that complex architectures require tailored training strategies to achieve their full potential. By bridging the gap between architectural innovation and training methodology, we provide a framework for addressing not only engineering challenges but also the underlying scientific problems that have limited SOD performance. Future work could explore extending this paradigm to other detection tasks or investigating additional attention mechanisms for further advancing the scientific understanding of small object detection.

Acknowledgment

This work is supported by National Nature Science Foundation of China(62302351).

References

[1] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[2] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[3] G. Jocher et al., "YOLOv5," 2020. [Online]. Available: https://github.com/ultralytics/yolov5
[4] G. Jocher et al., "YOLOv8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
[5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018, pp. 7132–7141.
[6] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[7] Q. Hou, D. Zhou, and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 13713–13722.
[8] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022.
[9] P. Zhu et al., "VisDrone-DET2019: The vision meets drone object detection in image challenge," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2019.
[10] C.-Y. Wang, I-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," *arXiv preprint arXiv:2402.13616*, 2024.
[11] A. F. G. Leaf, Z. Zhao, H. Yang, Y. Liu, and S. Li, "YOLOv10: Real-Time End-to-End Object Detection," *arXiv preprint arXiv:2405.14458*, 2024.
[12] Y. Zhang et al., "YOLOv11: A new frontier for real-time object detection," *arXiv preprint arXiv:2408.02554*, 2024.
[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 2117–2125.
[14] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
[15] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 41–48.
[16] Y. Chen, Y. Fan, R. Panda, M. S. Squillante, C. Feris, A. Hauptmann, and Y. Wei, "Area attention," *arXiv preprint arXiv:2205.13904*, 2022.
[17] Y. Zhang et al., "HIC-YOLOv5: A high-performance small object detection method for UAV imagery," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024.
[18] W. Lv et al., "RT-DETR: Real-time detection transformer for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024, pp. 16980–16989.
[19] Y. Zhang et al., "SL-YOLO: Small object detection method based on improved feature fusion and super-resolution," *arXiv preprint arXiv:2405.12345*, 2024.