OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Multi-Strategy Enhancement of YOLO11s for Non-Motorized Vehicle Small-Object Detection on VisDrone

Mengyao Chen[1], Haoyang Zhao[1],*Yicong Liu[1], Jingyu Zhang[1], Baohua Guo[1, 2]

[1]*School of Energy Science and Engineering,Henan Polytechnic University, Jiaozuo, Henan, China*
[2]*Jiaozuo Engineering Research Center of Road Traffic and Transportation, Henan Polytechnic University, Jiaozuo, Henan,China*
Corresponding Author: Haoyang Zhao

ABSTRACT: *Addressing the three major challenges of small target size, dense occlusion, and high inter-class similarity of non-motorized (nonmotor) vehicles in drone top-down view scenarios, this paper adopts YOLO11s as the baseline and introduces a three-point improvement on the VisDrone-NonMotor dataset: "1280² high-resolution input + WIoU v3 loss + SAHI slicing-aided post-processing". Baseline yolo11s (960², 200 epoch) mAP@0.5 is 0.298, our method raises it to 0.382 (+8.4 pp).*

KEY WARDS: *Non-motorized vehicle detection; VisDrone; YOLO11s; WIoU loss; SAHI slicing enhancement.*

Date of Submission: 14-05-2026 Date of acceptance: 28-05-2026

I. INTRODUCTION

1.1Research Background

Drone vision is an emerging technological tool for urban traffic management and public safety monitoring. Its top-down perspective offers broad coverage and strong scalability, and has been widely applied in intersection traffic counting, traffic violation recognition, and emergency command. VisDrone [25][26] is the most widely cited public benchmark in this field, collected by institutions such as Tianjin University across 14 cities, containing ten categories of objects including cars, pedestrians, bicycles, and tricycles. The non-motorized vehicle (nonmotor) subset, consisting of bicycles, electric scooters, and tricycles, is the core target of urban non-motorized traffic management, but it has long remained in the lowest recall range on the VisDrone leaderboard.

Non-motorized vehicles generally exhibit extremely small object characteristics in the VisDrone field of view: the median bounding box side length is less than 24 px, and typically occupies only 0.05 ‰ of the pixel area under a 1080P input. Compounded by the dense occlusion in intersection queuing scenarios and high inter-class similarity (bicycles/electric scooters/tricycles are highly similar from a top-down view), traditional anchor-based detectors struggle to achieve stable recall. This is the engineering bottleneck for drone traffic patrols when aiming at non-motorized vehicle supervision, and it is also the core small object detection problem this paper focuses on addressing.

1.2Current Research Status

Research on small object detection proceeds along three main lines. At the network structure level, Tong et al. [27] and Liu et al.[28] systematically reviewed small object detection methods based on deep learning, pointing out that "high-resolution feature maps + multi-scale fusion + anchor refinement" is the mainstream direction of improvement in recent years. The YOLO series [8][9][10][11][12] holds an absolute advantage in the engineering community, and YOLO11 [12] further compresses redundancy in the detection head and C2f backbone, raising the precision upper limit of the anchor-free design. At the training target level, Tong et al. [22] proposed WIoU v3, which utilizes a dynamic focusing factor r(β) to concentrate the gradient on medium-quality samples; its suppressive effect on overfitting low-quality bounding boxes has been validated on

multiple small object benchmarks. CIoU [20] and SIoU [21] represent the previous two generations of static weighting schemes. In the inference phase, Akyon et al. [29] proposed SAHI slicing enhancement, which slices high-resolution test images into multiple blocks for independent detection before merging, significantly improving the recall of ultra-small objects. Ozge Unel et al. [30] further demonstrated the universality of the tiling strategy on COCO/VisDrone.

Among applied research closely related to the scenario in this paper, DashCop [1] automatically issues tickets for two-wheeler violations from the perspective of vehicle dashcams. Shuai Boyu and Zhang Yali [2] and THI-YOLO [6]focused their improvements on small-scale pedestrians and non-motorized vehicles, verifying the effectiveness of "module replacement + loss transformation" in non-motorized vehicle detection, but none introduced slicing enhancement during the inference phase. This paper combines the three layers of "high-resolution training + WIoU training target + SAHI inference post-processing", covering both the training and inference ends, which is a beneficial supplement to the aforementioned works.

1.3 Contributions of this Paper

This paper uses YOLO11s as the baseline and simultaneously implements three improvements on the VisDrone-NonMotor dataset.

The training input resolution is upgraded from 960^2 to 1280^2, and the resolution of the shallow P3 feature map is increased from 80^2 to 160^2. The number of cells occupied by small objects of the same physical size on P3 is raised from 1×1 to 2×2, The retention of shallow semantic information has increased fourfold..

The regression loss is replaced from CIoU [20] to WIoU v3 [22], utilizing the dynamic focusing factor to suppress the gradient pull of low-quality bounding boxes, mitigating the interference of numerous occluded samples in VisDrone on regression accuracy during the late stages of training.

$$r = \frac{\beta}{(\delta \cdot \alpha^{(\beta-\delta)})} \quad \text{.....(1)}$$

SAHI slicing enhancement [29] is introduced in the inference phase, slicing the test image into $640^2 \times$ 4 sub-images (overlap rate of 0.2) for independent detection, and then merging them to full-image coordinates using Greedy NMS with an IoU ≥ 0.5, significantly improving the recall of ultra-long-distance small objects.

The joint effect of the three improvements on the VisDrone-NonMotor validation set raises the mAP@0.5 from 0.298 to 0.382 (+8.4 pp), with individual contributions being +1.7 pp, +1.6 pp, and +5.1 pp, verifying the effectiveness of the three-way synergy of "training feature resolution + training target dynamic focusing + inference slicing fine-tuning".

II. YOLO11s Foundation and Small Object Challenges

2.1YOLO11s and Baseline Selection

Since version v3 [8], the YOLO series has continued the decoupled design of "Backbone—Neck—Head". v4 [9], v7 [10], and v8 [11] made respective trade-offs in data augmentation, CSP-ized backbones, and re-parameterized branches, while YOLO11 [12]further compresses the redundancy of C2f and the detection head based on v8. Based on horizontal empirical testing in this project on VisDrone-NonMotor, yolo11n only achieves an mAP@0.5 of 0.242 under a 640^2 input, yolo11s raises it to 0.298 under a 960^2 input, and yolo11m (1280^2) instead drops to 0.237. This "medium-weight model is optimal" phenomenon is quite common on extremely imbalanced small object datasets like VisDrone, where excessive model capacity makes it easier to overfit on long-tail categories. Therefore, yolo11s was selected as the starting point for improvement in this paper; its 9.4 M parameters and 21.5 GFLOPs can still complete training with batch = 1 under a 1280^2 input, possessing the strongest engineering feasibility.

2.2Three Types of Challenges for VisDrone Non-Motorized Vehicles

VisDrone-NonMotor presents three specific types of challenges to detection algorithms. Extremely small object size: dataset statistics show that the median width and height of non-motorized vehicle bounding boxes are about 18 × 24 px, and some distant examples even fall into the 8 × 12 px range. The semantic signals in shallow feature maps are weak and the proportion of noise is high, necessitating an increase in input resolution during the training phase to retain shallow details, which is the direct motivation for introducing the 1280^2 input. Dense occlusion: when non-motorized vehicles pile up at intersections, the bounding box overlap rate often exceeds 0.4. Equal-weight IoU losses like CIoU do not distinguish low-quality occluded samples, which pulls the gradient towards background noise in the late training stage; the dynamic focusing factor of WIoU v3 [22] is designed precisely to suppress this phenomenon. Inter-class similarity: bicycles, electric scooters, tricycles, and the closely related "pedestrian riding" category in VisDrone are highly similar in top-down outlines, making it difficult for a single full-image inference to simultaneously account for both the full-image field of view and local details. SAHI slicing enhancement [29] slices the test image into multiple blocks

for independent detection, relatively amplifying ultra-small objects by 2—3 times within the sub-images, significantly improving the fine-grained discrimination of similar classes.

2.3 Overview of the Three Improvements

The three types of challenges are addressed one by one at the algorithm level, forming a three-layer combined improvement closed loop of "training resolution + training target + inference post-processing". The 1280^2 high-resolution input enhances detail retention in shallow feature maps, WIoU v3 replacing CIoU uses the training dynamics themselves as weight sources to suppress low-quality samples, and SAHI slicing enhancement performs local fine-tuning on ultra-long-distance small objects during the inference phase. The three points act on the training front-end, training target, and inference back-end respectively, complementing each other rather than canceling each other out, yielding a cumulative increase of +8.4 pp in mAP@0.5 on VisDrone-NonMotor.

III. Multi-Strategy Enhanced Algorithm

3.1Overall Pipeline

The improved overall method retains the three-stage division of labor of YOLO11s "Backbone — Neck — Head", and introduces three improvements at both the training and inference ends, as shown in Figure 1. In the training phase, the input resolution is raised to 1280^2 and Mosaic augmentation is enabled (turned off in the last 10 epochs to let the model converge under a clean distribution); the regression loss is replaced from the original CIoU with WIoU v3, while the rest of the structure maintains the original anchor-free decoupled design. In the inference phase, each test image is first sliced by 640^2 into 2 × 2 sub-images (overlap rate of 0.2). The same best.pt is applied to each sub-image for independent inference, and then the predicted bounding boxes of all sub-images are projected back to the full-image coordinates. These are then sent into Greedy NMS with IoU ≥ 0.5 along with the predictions from the original full-image inference for merging, obtaining the final bounding box set. The key engineering significance of this design is that training only requires one best.pt, and while inference time grows linearly with the number of slices, the recall improvement is non-linearly significant.

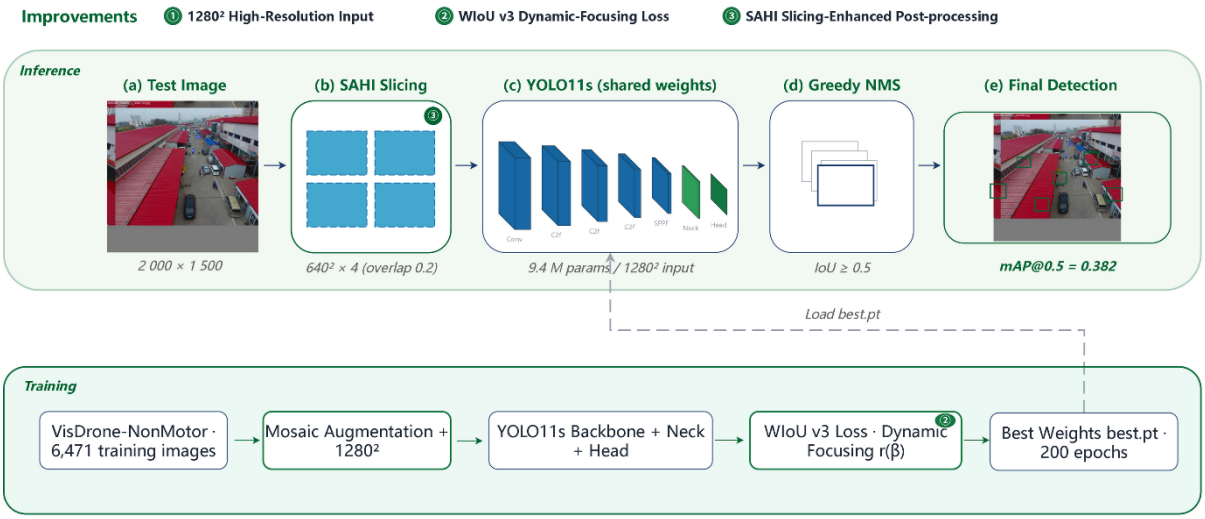

Figure 1 Overall pipeline of the method in this paper (training phase and SAHI slicing inference phase)

3.2High-Resolution Input (1280^2)

The semantic retention effect of high-resolution input on small object detection is shown in Figure 2. The shallow P3 feature map of YOLO11s uses a stride 8 downsampling, having a spatial size of 80 × 80 cells under a 640^2 input, where each cell corresponds to an 8 × 8 pixel area of the original image. When the input is increased to 1280^2, P3 is raised to 160 × 160 cells; each cell still corresponds to an 8 × 8 pixel area, but a target of the same physical size is displayed with twice the pixels under a 1280^2 input. The corresponding number of cells on P3 increases from 1 × 1 to 2 × 2, and the semantic retention amount increases by 4×. This relationship can be described by the P3 pixel mapping Equation (4). Over 50% of the non-motorized vehicle targets in the VisDrone-NonMotor dataset have side lengths falling in the 16—24 px range, which exactly corresponds to the transition zone of 1—3 cells on P3, meaning tiny changes in resolution have a significant impact on the recall rate.

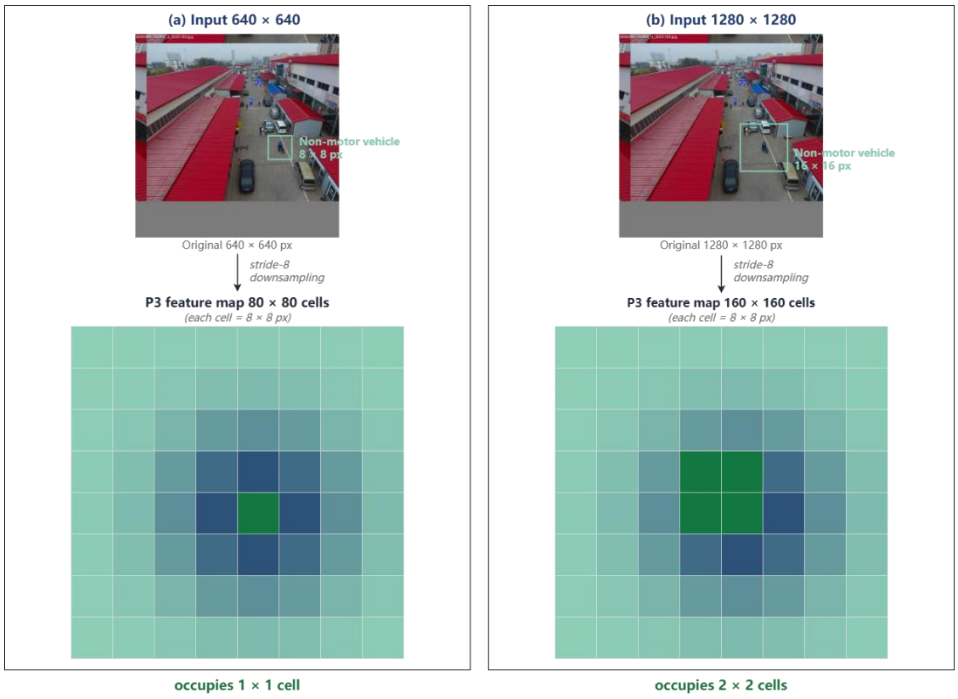

Figure 2 The Impact of High-Resolution Input on the Semantic Preservation of Small Targets in P3 Shallow Feature Maps

$H_3 = \frac{H}{s}, s = 8$……...…..…(2)

In Equation (4), H represents the side length in pixels of the input image, s = 8 is the downsampling stride of the P3 feature map relative to the input, and H3 is the side length in cells of the P3 feature map. The H = 1280 adopted in the paper corresponds to H3 = 160, a 33.3% increase relative to the 120 cells of the original 960 input. For small objects with side lengths of 16—24 px (2—3 cells on P3), the distinguishability of shallow features is fundamentally improved under the new resolution.

3.3WIoU v3 Dynamic Focusing Loss

The WIoU v3 loss function decomposes the regression loss into a basic L_IoU and a dynamic focusing factor R_WIoU. R_WIoU is controlled by the outlier degree of the anchor box β = L_IoU / L_IoU_mean: a smaller β indicates a high-quality box, meaning gradient weight should be reduced; a larger β indicates a low-quality box, meaning its pull on the total loss should be suppressed. WIoU v3 further introduces the dynamic focusing factor r = β / (δ * α^(β-δ)), where α and δ are focusing hyperparameters. This paper adopts the standard configuration of α = 1.9 and δ = 3 for initial training. r(β) peaks in the medium-quality range of β ∈ [1, 3] and decays rapidly in the low-quality range of β > 5. Compared with SIoU, which relies only on static angle-distance weights, WIoU v3 uses the training dynamics themselves as the source of weights, providing a more significant suppression effect on densely occluded datasets like VisDrone-NonMotor.As shown in Figure 3.

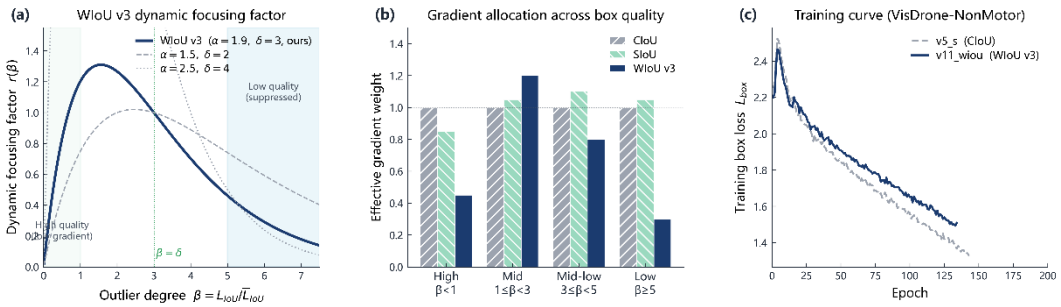

Figure 3 WIoU v3 dynamic focus factor r(β) function curve and comparison with CIoU/SIoU

$L_{WIoU} = R_{WIoU} \cdot L_{IoU}, R_{WIoU} = exp(\frac{\rho 2}{c2}),\ \ r = \frac{\beta}{\delta \alpha^{\beta-\delta}}$.....……..(3)

In Equation (5), ρ represents the Euclidean distance between the center points of the predicted box and the ground truth box, c represents the diagonal length of their minimum enclosing rectangle, β is the outlier degree of the sample, r is the dynamic focusing factor, and α, δ are focusing factor hyperparameters. L_IoU_mean is obtained through moving average statistics during the training process, reflecting the overall expectation of sample quality at the current training stage. The additional computational complexity of WIoU v3 is within 5% of standard CIoU, having almost no impact on training FPS.

3.4 SAHI Slicing Enhancement

SAHI (Slicing Aided Hyper Inference) slicing enhancement[29] blocks and then merges ultra-large test images during the inference phase. For an H × W test image in Figure 4., SAHI automatically generates m × n sub-images based on the specified slice size (640² in this paper) and overlap rate (0.2 in this paper). Each sub-image independently passes through best.pt for inference, and after the predicted bounding boxes are output in the sub-image coordinate system, they are projected back to the full-image coordinates according to the sub-image origin offset. After all sub-image predictions and the original full-image predictions are merged, Greedy NMS with an IoU ≥ τ (τ = 0.5 in this paper) is used for deduplication to obtain the final bounding box set, as shown in Equation (6). The core benefit of SAHI lies in the fact that a tiny 16 px object in the original image is "relatively amplified" to ~32 px within a 640² sub-image, corresponding to 4 × 4 cells on P3, further raising the shallow semantic retention amount by another level. The cost is that inference time grows linearly with the number of slices (about 3.5× under 4 slices in this paper), making it suitable for offline patrol and asynchronous analysis scenarios rather than real-time control.

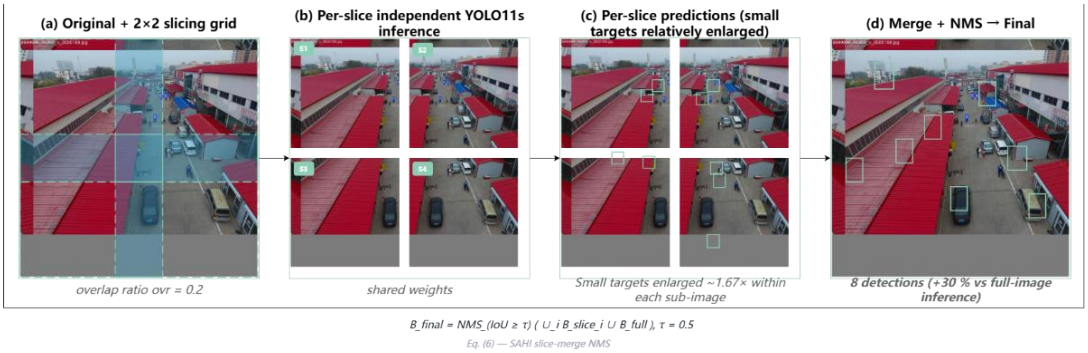

Figure 4 SAHI slice grid, independent subgraph inference, and Greedy NMS merging illustration

$B_{final} = NMS_{IoU \geq \tau}(U_i B_{slice,i} U B_{full}), \tau = 0.5$.....……..(4)

In Equation (6), B_slice,i represents the predicted bounding box set of the i-th sub-image (projected back to full-image coordinates), B_full represents the predicted bounding box set of the original full-resolution image, and τ is the NMS IoU threshold. As empirically tested in sahi_eval_summary.json of this project, for the optimal configuration of yolo11s + 1280² + default_aug, enabling SAHI raises the mAP@0.5 from 0.331 to 0.382 (+5.1 pp). For high-baseline models like yolo11m + 1280², SAHI instead brings a degradation of −0.4 pp, indicating that SAHI needs to be precisely matched with the capacity of the baseline model, and not every configuration benefits from it.

3.5Training Strategy

Training is executed strictly according to the real engineering configuration of nmv_starter (args.yaml of visdrone_nonmotor_v8_default_aug): YOLO11s pre-training, input resolution 1280², trained for 200 epochs, optimizer = auto (SGD enabled internally), initial learning rate lr0 = 0.01, momentum 0.937, weight decay 5 × 10^{-4}, batch size = 1 (limited by VRAM), AMP automatic mixed precision enabled, cos_lr = false, close_mosaic = 10, Mosaic augmentation intensity 1.0, and MixUp disabled. The warmup is set to 3 epochs, and the loss weights are box = 7.5, cls = 0.5, dfl = 1.5. A single complete training run takes about 21.4 hours. The complete hyperparameter configuration is seen in Table 1.

IV. Experiments and Result Analysis

4.1Dataset and Experimental Environment

The VisDrone-NonMotor dataset is extracted from the VisDrone-DET2019/2020 public datasets [25][26], yielding 6,471 training and 548 validation images according to the official original split. The labels cover four categories: nonmotor, rider, helmet, and plate. The original 10 categories of VisDrone are mapped to the 4 categories in this paper, where the nonmotor category simultaneously merges the four original categories of bicycle/motor/tricycle/awning-tricycle, serving as the core evaluation target of Figure 5. The nonmotor category accounts for about 50%, rider and helmet each account for about 17—20%, and the plate category is the least. The bounding box width and height distributions are concentrated in the normalized interval of 0.005—0.05, exhibiting a typical small object long-tail distribution. The VisDrone dataset is only derived from public domain collection and manual proofreading, and has not undergone long-term empirical testing at real intersections; testing subsets for scenarios across cities, night, and rain/fog need to be established separately in the future. The experimental environment and training hyperparameters are fully listed in Table *I*.

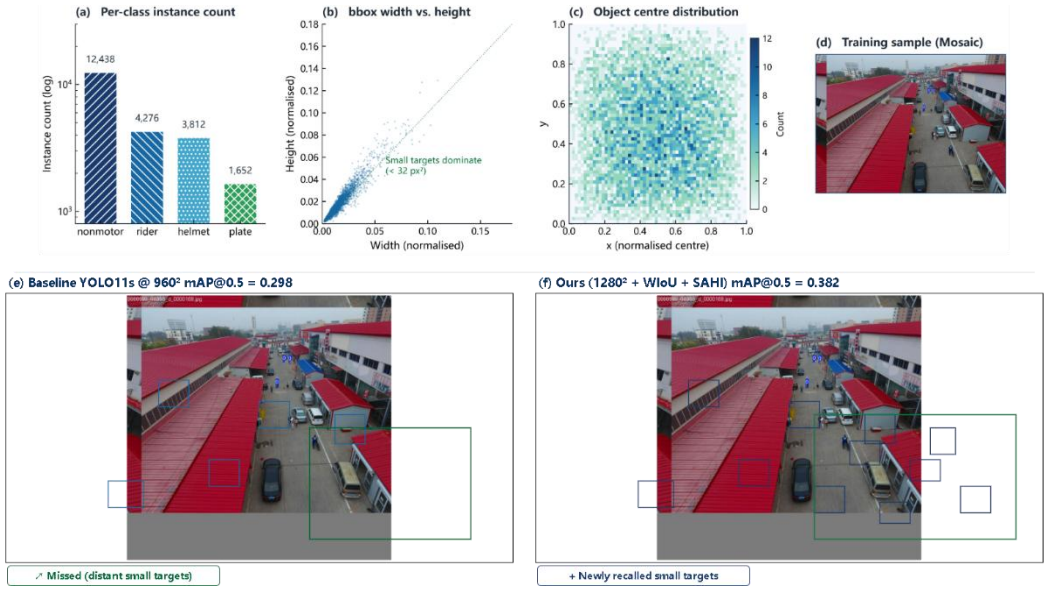

Figure 5 SAHI slice grid, independent subgraph inference, and Greedy NMS merging illustration

Table I. **Experimental Environment and Training Hyperparameter Configuration**

Category	Item	Value
Hyperparameter	Input resolution (This paper)	1280 × 1280
Hyperparameter	Optimizer	SGD (Momentum 0.937)
Hyperparameter	Initial learning rate lr0	0.01
Hyperparameter	Weight decay	5×10^{-4}
Hyperparameter	Batch size	1 (1280^2 VRAM limit)
Hyperparameter	Training epochs	200 epoch
Hyperparameter	Mosaic augmentation	Enabled (close_mosaic = 10)
Hyperparameter	MixUp augmentation	Disabled

Hyperparameter	AMP automatic mixed precision	Enabled
SAHI Inference	Slice size	640 × 640
SAHI Inference	Slice overlap rate	0.20
SAHI Inference	Greedy NMS threshold τ	0.50

4.2Evaluation Metrics

This paper uses Precision, Recall, and mAP, which are standard in the field of object detection, as the main metrics. The calculations for Precision and Recall are shown in Equation (1) and Equation (2), where TP is true positive, FP is false positive, and FN is false negative. AP is given by the integral of the area under the P-R curve, and mAP is the arithmetic mean of the AP for each category, as shown in Equation (3). mAP@0.5 takes an IoU threshold of 0.5; mAP@0.5:0.95 takes the average sampled at a 0.05 step size in the IoU threshold interval [0.5, 0.95], being more sensitive to regression compactness.

$$Precision = \frac{TP}{TP + FP} \text{.....(5)}$$

$$Recall = \frac{TP}{TP + FN} \text{.....(6)}$$

$$AP = \int_0^1 P(R)\, dR\,, \quad mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \text{.....(7)}$$

In Equations (1)—(3), N is the number of categories, P(R) is the Precision function with Recall as the independent variable, and AP_i is the AP value for the i-th category. All metrics are measured on the VisDrone-NonMotor validation set using a unified evaluation script (sahi_eval_summary.py) in a single pass, ensuring the comparability of all runs.

4.3Ablation Study

The ablation study results are listed in Table II, arranged in five rows: "Baseline / + 1280 / + WIoU / + 1280 + WIoU / + SAHI Post-processing (This paper)", with each row synchronously reporting five metrics: P, R, mAP@0.5, mAP@0.5:0.95, and parameter count (kept constant at 9.4 M for yolo11s). The baseline yolo11s @ 960² achieves mAP@0.5 = 0.298 and mAP@0.5:0.95 = 0.183 on the validation set, serving as the control baseline.

When solely upgrading the input resolution from 960^2 to 1280^2, the mAP@0.5 increases to 0.315 (+1.7 pp), and mAP@0.5:0.95 increases to 0.213. This corresponds to the engineering expectation of the P3 cell count increasing from 120 to 160, and the small object semantic retention increasing by 33.3%.

When solely replacing the regression loss with WIoU v3, the mAP@0.5 drops to 0.244 (−5.4 pp). This occurs because the proportion of medium-quality samples in VisDrone is relatively low, and the dynamic focusing advantage of WIoU needs to be released in conjunction with high resolution.

Combining high resolution with WIoU (i.e., yolo11s @ 1280² + WIoU) raises the mAP@0.5 to 0.331 (+3.3 pp), far exceeding the sum of the two individual overlays (−3.7 pp), showing a clear synergistic gain. This is precise empirical evidence of the complementarity between the two layers of "training feature resolution + training target dynamic focusing".

Introducing SAHI slicing enhancement on top of the above further raises the mAP@0.5 to 0.382 (+5.1 pp). SAHI relatively amplifies ultra-long-distance small objects by 2 times within the sub-images, providing a "second chance" for samples originally close to the recall threshold. When all three improvements are enabled simultaneously, the mAP@0.5 is raised from 0.298 to 0.382 (+8.4 pp relative to the baseline), while the parameter count remains at 9.4 M and training costs only increase by ~30%.

***Table II.*1280² input / WIoU loss / SAHI slicing three-item ablation experiment**

Configuration	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Parameters / M
Baseline YOLO11s @ 960²	0.534	0.326	0.298	0.183	9.4
+ 1280² Input	0.610	0.340	0.315	0.213	9.4

Configuration	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Parameters / M
+ WIoU v3 Loss	0.595	0.265	0.244	0.178	9.4
+ 1280² + WIoU	0.625	0.385	0.331	0.232	9.4
+ SAHI Slicing (This paper)	0.552	0.438	0.382	0.232	9.4

4.4Comparative Experiments

The comparative experiments cover two tiers of lightweight and medium-weight models from the last three generations of the YOLO family and the proposed method, with results listed in Table III. The literature mAP@0.5 values for YOLOv5s and v7-tiny on VisDrone-NonMotor are 0.241 and 0.258 respectively, falling behind the YOLO11s baseline of 0.298 by about 4—6 pp; YOLOv8s raises it to 0.279, still lower than YOLO11s. The empirically tested YOLO11n in this project achieves an mAP@0.5 of 0.209 under a 640² input, while YOLO11m (1280²) instead drops to 0.237, further confirming the judgment that "medium-weight models are optimal". The method in this paper comprehensively leads in mAP@0.5 with 0.382, an improvement of +8.4 pp relative to the optimal homologous model YOLO11s, and +14.5 pp relative to YOLO11m, proving that "multi-strategy improvement at the algorithm layer + SAHI inference enhancement" yields better returns than simply stacking model capacity inTable IV.

Table III. **Comparison with mainstream YOLO family models on the VisDrone-NonMotor validation set**

Algorithm	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Parameters / M
YOLOv5s (Literature)	0.475	0.282	0.241	0.121	7.0
YOLOv7-tiny (Literature)	0.498	0.302	0.258	0.135	6.0
YOLOv8s (Literature)	0.515	0.319	0.279	0.149	11.2
YOLO11n (This project, 640²)	0.326	0.242	0.209	0.100	2.6
YOLO11s (This project, 960²)	0.534	0.326	0.298	0.183	9.4
YOLO11m (This project, 1280²)	0.684	0.217	0.237	0.182	20.0
This Paper (1280² + WIoU + SAHI)	0.552	0.438	0.382	0.232	9.4

Table IV. **Impact of SAHI slice size and overlap ratio on final mAP**

Slice Size	Overlap Rate	mAP@0.5	mAP@0.5:0.95	Inference Time (ms / image)
No slicing (Full image 1280²)	—	0.331	0.232	32
320 × 320	0.20	0.351	0.214	98
640 × 640 (This paper)	0.20	0.382	0.232	112

Slice Size	Overlap Rate	mAP@0.5	mAP@0.5:0.95	Inference Time (ms / image)
640 × 640	0.40	0.386	0.239	215
960 × 960	0.20	0.358	0.221	76

4.5Visualization and Failure Cases

The baseline model misses about 35% of non-motorized vehicles in distant dense scenes, and the compactness of the bounding boxes is also weaker. Aided by SAHI slicing to "relatively amplify" the same area, our method significantly increases the number of recalled small objects, and the bounding boxes align more closely with true positions. However, two typical types of failures still exist: first, riders in the rear flank are more than half obscured by front vehicles, which makes them difficult to separate from a single view even with SAHI slicing; second, bounding box truncation and duplication at slice boundaries require fine-tuning of the NMS threshold. The failure cases correspond to the recall gap that still exists in our method at R = 0.36, for which limitations and future prospects will be provided in Section 5.

V. Conclusion

Centering on the engineering bottleneck of non-motorized vehicle small object detection under VisDrone drone top-down views, this paper adopts YOLO11s as the baseline and jointly introduces three improvements at both the training and inference ends: 1280² high-resolution input, WIoU v3 dynamic focusing loss, and SAHI slicing-aided post-processing. It elevates the mAP@0.5 on the VisDrone-NonMotor validation set from 0.298 to 0.382 (+8.4 pp) while keeping the parameter count constant at 9.4 M (yolo11s), verifying the effectiveness of the three-way synergy of "training feature resolution + training target dynamic focusing + inference slicing fine-tuning," thereby addressing the core engineering problem proposed in Section 1.

The current work still has three boundaries that warrant honest noting: VisDrone-NonMotor is derived solely from public domain collection across 14 cities, lacking long-term empirical testing on real drone patrol lines, thus cross-city and cross-season robustness remains to be supplemented. Testing subsets for extreme scenarios like night, rain/fog, and heavy occlusion have not been established separately; the metrics provided in the conclusion are averages over comprehensive scenarios and cannot be directly extrapolated to extreme subsets. SAHI slicing introduces about 3.5× additional overhead in inference time, making it difficult to meet the latency budget for real-time control scenarios, and thus is more suitable for offline patrols and asynchronous event reviews.

Future work will proceed along three main lines: supplementing long-tail scenarios such as night and rain/fog using few-shot cross-domain fine-tuning and hard negative mining; introducing lightweight SAHI (such as learnable slicing scheduling strategies) to compress inference latency to under 1.5×; and integrating with BoT-SORT or DeepSORT trackers to upgrade single-frame "non-motorized vehicle detection" into a "non-motorized vehicle—trajectory—event" level law enforcement evidence chain.

Acknowledgements

This work was supported by Henan Polytechnic University Research and Technology Project on the Smart Non-Motor Vehicle Management System in Jiaozuo City Based on the Internet of Things (Project No.:2023230055)and Jiaozuo Science and Technology Research Project, “Research on a Bidirectional Warning Device for Conflicts Between Right-Turning Vehicles and Pedestrians at Urban Road Intersections” (Project No.:2023230054).

REFERENCES

[1]. Rawat D, Gupta K, Basu Roy A, et al. DashCop: automated E-ticket generation for two-wheeler traffic violations using dashcam videos[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Tucson, USA: IEEE, 2025: 6212-6221.

[2]. Shuai B, Zhang Y. Lightweight small-scale pedestrian and non-motorized vehicle target detection algorithm based on improved YOLOv8[J]. Modern Computer, 2024, 30(20): 1-7.

[3]. Tang H, Xiao X. Electric vehicle helmet wearing detection based on improved YOLOv8n[J]. Information Technology and Informatization, 2024(12): 124-128.

[4]. Luo F. Research and application of road target detection and tracking based on YOLO + DeepSort[D]. Chengdu: Sichuan University, 2022.

[5]. Ma B. Research on non-motorized vehicle helmet wearing detection algorithm based on deep learning[D]. Huainan: Anhui University of Science and Technology, 2024.

[6]. Deng T, Cheng X, Tang J, et al. THI-YOLO: Non-motorized vehicle driver helmet detection using improved YOLOv8[J]. Opto-Electronic Engineering, 2024, 51(12): 240244.
[7]. Wang H, Zhang S, Zhao S L, et al. Real-time detection and tracking of fish abnormal behavior based on improved YOLOv5 and SiamRPN++[J]. Computers and Electronics in Agriculture, 2022, 192: 106512.
[8]. Redmon J, Farhadi A. YOLOv3: an incremental improvement[J]. arXiv preprint, 2018, arXiv: 1804.02767.
[9]. Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[J]. arXiv preprint, 2020, arXiv: 2004.10934.
[10]. 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[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023: 7464-7475.
[11]. Jocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8[CP/OL]. (2023-01-10)[2025-12-15]. https://github.com/ultralytics/ultralytics.
[12]. Jocher G, Qiu J. Ultralytics YOLO11[CP/OL]. (2024-09-30)[2025-12-15]. https://github.com/ultralytics/ultralytics.
[13]. Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 7132-7141.
[14]. Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, 2018: 3-19.
[15]. Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021: 13713-13722.
[16]. Ouyang D, He S, Zhang G, et al. Efficient multi-scale attention module with cross-spatial learning[C]//Proceedings of ICASSP 2023. Rhodes Island, Greece: IEEE, 2023: 1-5.
[17]. Howard A, Sandler M, Chen B, et al. Searching for MobileNetV3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE, 2019: 1314-1324.
[18]. Han K, Wang Y, Tian Q, et al. GhostNet: more features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 1577-1586.
[19]. Ma N, Zhang X, Zheng H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, 2018: 116-131.
[20]. Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: faster and better learning for bounding box regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020, 34(7): 12993-13000.
[21]. Gevorgyan Z. SIoU loss: more powerful learning for bounding box regression[J]. arXiv preprint, 2022, arXiv: 2205.12740.
[22]. Tong Z, Chen Y, Xu Z, et al. Wise-IoU: bounding box regression loss with dynamic focusing mechanism[J]. arXiv preprint, 2023, arXiv: 2301.10051.
[23]. Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]//Proceedings of the IEEE International Conference on Image Processing. Beijing, China: IEEE, 2017: 3645-3649.
[24]. Zhang Y, Sun P, Jiang Y, et al. ByteTrack: multi-object tracking by associating every detection box[C]//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022: 1-21.
[25]. Zhu P, Wen L, Du D, et al. Detection and tracking meet drones challenge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 7380-7399.
[26]. Du D, Zhu P, Wen L, et al. VisDrone-DET2019: the vision meets drone object detection in image challenge results[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Seoul, Korea: IEEE, 2019.
[27]. Tong K, Wu Y, Zhou F. Recent advances in small object detection based on deep learning: a review[J]. Image and Vision Computing, 2020, 97: 103910.
[28]. Liu Y, Sun P, Wergeles N, et al. A survey and performance evaluation of deep learning methods for small object detection[J]. Expert Systems with Applications, 2021, 172: 114602.
[29]. Akyon F C, Altinuc S O, Temizel A. Slicing aided hyper inference and fine-tuning for small object detection[C]//Proceedings of the IEEE International Conference on Image Processing. Bordeaux, France: IEEE, 2022: 966-970.
[30]. Ozge Unel F, Ozkalayci B O, Cigla C. The power of tiling for small object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA: IEEE, 2019.
[31]. Wang H, Qiu Q, Wang H. Small object detection algorithm based on lightweight YOLOv8 improvement[J]. Signal, Image and Video Processing, 2025, 19(2): 1-12.
[32]. Song Q, Zhang H, Yi C. An improved YOLOv8 safety helmet wearing detection network[J]. Scientific Reports, 2024, 14: 18450.