ISSN: 1813-4890

MID-YOLO11: A Lightweight UAV Small-Target Pedestrian Detection Method

Siyu Jiang

School of Computer Science, Xi'an Shiyou University, Xi'an 710000, China

Abstract

This paper proposes MID-YOLO11, a lightweight and efficient object detection model for small-target pedestrian detection in UAV aerial photography scenarios. Based on YOLO11n, three core improvements are introduced: (1) MSFHA (Multi-Scale Spatial-Frequency Hybrid Attention), which achieves hybrid spatial-frequency feature enhancement through multi-scale context aggregation and frequency-aware gating; (2) ISCRConv (Integrated Spatial-Channel Reconstruction Convolution), which eliminates feature redundancy and preserves small-target edge details while supporting optional downsampling; and (3) DTGHDetect (Dual-Gated Task-Guided Head), which achieves precise task alignment through decoupled branches and dual-gated task interaction. On the preprocessed VisDrone2019 pedestrian detection subset, MID-YOLO11 achieves mAP@0.5 of 45.0%, which is 2.6 percentage points higher than the baseline YOLO11n, with parameters reduced to 2.45M and GFLOPs of only 6.5, achieving an excellent balance between accuracy and efficiency. Ablation experiments verify the effectiveness and synergistic gain of each module, and comparison experiments show that the proposed method comprehensively outperforms mainstream lightweight models such as YOLOv8n, YOLOv10n, and YOLO12n under comparable parameter counts.

Keywords

UAV object detection; small-target pedestrian; attention mechanism; feature reconstruction; task alignment.

1. Introduction

UAV (Unmanned Aerial Vehicle) low-altitude remote sensing technology has been widely applied in energy infrastructure inspection and security surveillance due to its flexibility and wide coverage. In large industrial areas such as oilfield photovoltaic power stations, using UAVs for high-frequency patrol and timely detection of intruders is an important means to reduce safety risks and ensure stable facility operation. However, pedestrian detection in UAV aerial photography faces unique challenges: extremely small target scales (usually less than 0.02% of the image area), complex background textures (regular geometric patterns of photovoltaic panel arrays and ground reflections), large variation in target density, and strict real-time processing requirements, making it difficult for generic object detection methods to directly meet application needs.

The YOLO (You Only Look Once) series has occupied an important position in real-time object detection due to its end-to-end detection and fast inference speed [1-5]. However, existing YOLO lightweight models still have obvious shortcomings in small-target pedestrian detection under high-altitude overhead scenarios: (1) severe loss of small-target information in deep features; (2) insufficient utilization of frequency-domain information by attention mechanisms; (3) imperfect task alignment mechanisms in detection heads, resulting in poor consistency between localization and classification. To address these problems, this paper proposes MID-YOLO11 (Multi-Scale Integrated Detection YOLO11). The main contributions are: (1) MSFHA

ISSN: 1813-4890

attention module combining dual-branch parallel grouped convolution with a learnable Laplacian high-pass filter for multi-scale spatial-frequency hybrid feature enhancement; (2) ISCRConv reconstruction convolution that reduces parameters while preserving small-target details through optional downsampling and spatial-channel dual reconstruction; (3) DTGHDetect detection head that improves the synergistic accuracy of small-target classification and localization through decoupled asymmetric branches and the Dual-Gated Task Interaction (DGTI) module.

2. Related Work

2.1. Object Detection in UAV Scenarios

For object detection from UAV perspectives, existing methods mainly focus on the following directions. In multi-scale feature fusion, FPN and PANet propagate semantic information through top-down and bottom-up bidirectional paths, effectively improving detection accuracy for multi-scale targets [6]. However, the problem of detail loss in deep features for extremely small-scale targets has not been completely resolved. In attention mechanisms, SE, CBAM, and SimAM enhance model response to discriminative features by reweighting feature channels or spatial positions [7]. Nevertheless, existing attention mechanisms are mostly limited to the spatial or channel domain with limited frequency-domain utilization. In lightweight detection head design, decoupled detection heads separate localization and classification, and TAL jointly optimizes localization and classification scores, but its task interaction mechanism is relatively simple with insufficient channel-level dependency exploration.

2.2. Small-Target Detection Methods

Small-target detection is a long-standing challenge in computer vision. Super-resolution reconstruction methods improve target resolution through generative networks but incur large computational overhead. Data augmentation methods such as Mosaic and Copy-Paste alleviate small-target scarcity in training sets. Recently, frequency-domain analysis methods have been introduced to extract high-frequency edge information for small-target recognition, but their lightweight integration with mainstream detection frameworks remains an open problem. The MSFHA module proposed in this paper integrates frequency-domain gating with multi-scale attention into a lightweight unified module, filling this research gap.

3. MID-YOLO11 Architecture

3.1. Overall Network Structure

MID-YOLO11 follows the classic three-stage architecture of YOLO11 — Backbone, Feature Pyramid Neck, and Detection Head — with the proposed modules introduced at key positions. As shown in Fig. 1, MSFHA attention modules are embedded after the P3, P4, and P5 feature layers of the backbone. Downsampling convolutions at P4 and P5 stages are replaced with ISCRConv to reduce high-frequency detail loss. DTGHDetect replaces the original detection head, outputting three-scale predictions at P3/8, P4/16, and P5/32 for small, medium, and large targets respectively.

ISSN: 1813-4890

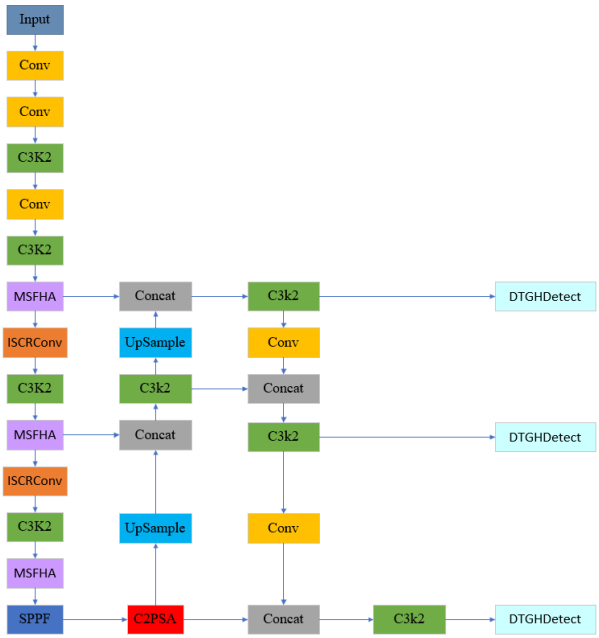

Fig. 1. MID-YOLO11 overall network architecture

3.2. MSFHA Attention Mechanism

The original SimAM computes pixel-level energy based on statistical features for lightweight attention, but has two limitations: insufficient multi-scale context capture and absent frequency-domain information utilization, leading to poor small-target performance. To address this, MSFHA (Multi-Scale Spatial-Frequency Hybrid Attention) integrates four core modules: multi-scale context aggregation, frequency-aware gating, statistical energy computation, and residual energy refinement, as illustrated in Fig. 2.

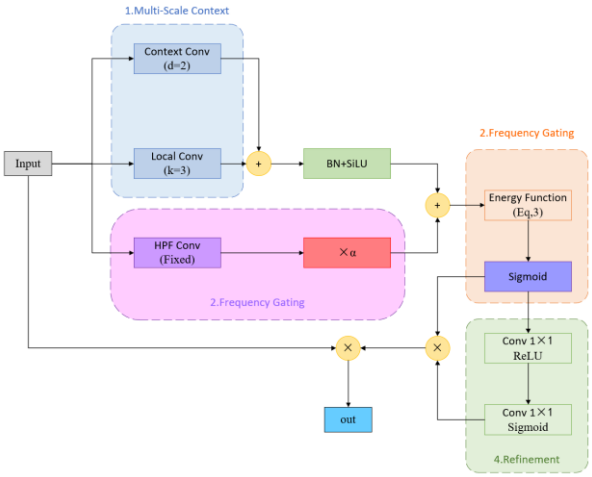

Fig. 2. MSFHA attention mechanism structure

For multi-scale context aggregation, MSFHA adopts a dual-branch parallel grouped convolution structure. The local branch uses a 3×3 standard depth-grouped convolution (stride=1, padding=1) to extract fine-grained local details critical for small-target recognition; the global branch uses a 3×3 dilated depth-grouped convolution (dilation=2, padding=2) to extend the receptive field to an effective 5×5 for capturing long-range spatial dependencies. Their outputs are element-wise added, then processed through BN and SiLU activation.

For frequency-aware gating, MSFHA designs a learnable Laplacian high-pass filter (HPF) implemented as a 3×3 depth-grouped convolution with Laplacian-initialized kernels, effectively extracting high-frequency edge and texture components. A channel-wise learnable parameter α (initialized to 0.5) dynamically balances high-frequency and spatial features. The

ISSN: 1813-4890

statistical energy computation and residual refinement follow the original SimAM approach, using a lightweight residual bottleneck to refine the attention map before element-wise multiplication with the input.

3.3. ISCRConv

The original ScConv achieves feature redundancy elimination through serial SRU (Spatial Reconstruction Unit) and CRU (Channel Reconstruction Unit), but only supports stride=1 processing [8]. ISCRConv (Integrated Spatial-Channel Reconstruction Convolution) adds an optional downsampling function, as shown in Fig. 3. The optional downsampling module switches adaptively: when stride s>1, it uses "average pooling + 1×1 convolution + BN"; when s=1, identity mapping is used, consistent with the original ScConv.

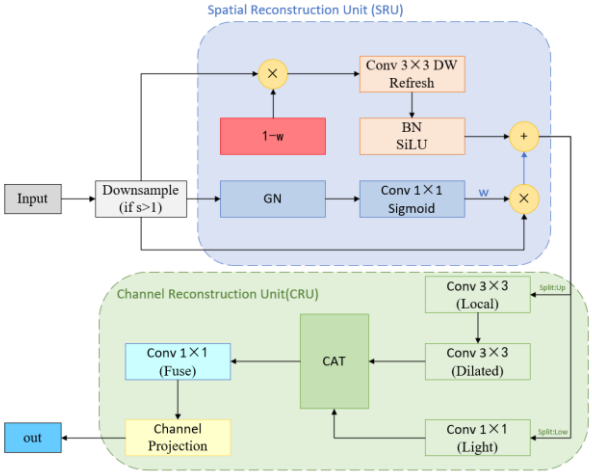

Fig. 3. ISCRConv module structure

SRU eliminates spatial feature redundancy by generating spatial attention weight maps and balancing feature retention and refresh through weight-preserving and depth convolution refresh branches. CRU then processes the SRU output with three parallel convolution branches: a local branch (3×3 standard conv), a dilated branch (3×3 dilated conv, dilation=2), and a lightweight branch (1×1 conv), capturing local spatial correlations, long-range channel dependencies, and linear channel relationships respectively. Their concatenated outputs are processed by a fusion convolution for dimensionality reduction.

3.4. DTGHDetect Detection Head

The original TAL detection head uses simple spatial multiplication for branch interaction, with insufficient channel-level dependency exploration and lacking targeted optimization for small targets. DTGHDetect (Dual-Gated Task-Guided Head) introduces three core improvements: MSFHA input enhancement, DGTI (Dual-Gated Task Interaction), and decoupled branch feature extraction, as shown in Fig. 4.

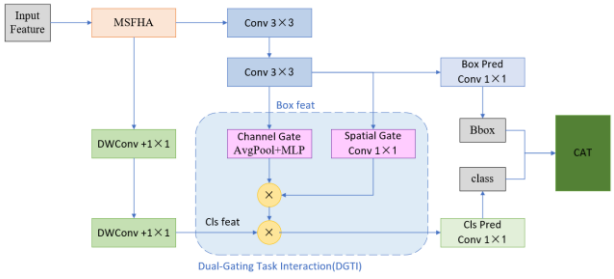

Fig. 4. DTGHDetect detection head structure

Multi-scale input features are first enhanced by MSFHA, then split into decoupled Box (localization) and Cls (classification) branches. The Box branch uses two serially connected 3×3 standard convolutions for high-capacity spatial localization features. The Cls branch uses two

ISSN: 1813-4890

serially connected lightweight (3×3 depthwise + 1×1 pointwise) convolutions. The DGTI module uses Box branch features to guide spatial and channel dual-gated alignment of Cls branch features with residual connections, achieving precise task alignment and effectively improving small-target classification accuracy.

4. Experiments and Results

4.1. Dataset

This study uses the VisDrone2019 dataset, released by the AISKYEYE team of Tianjin University in 2019, one of the most representative public UAV aerial photography datasets [9]. It covers 14 cities in China, containing 10,209 high-resolution static images, 261,908 video frames from 288 video clips, and over 2.6 million annotated bounding boxes across 10 classes. Targeted preprocessing is applied for pedestrian detection: (1) merging "pedestrian" and "people" classes into "person"; (2) removing non-"person" annotations; (3) filtering images without pedestrian targets; (4) retaining only samples with target scale smaller than 50×50 pixels. The resulting dataset split is shown in Table 1.

Table 1. VisDrone2019 pedestrian detection subset

Split	Images	Annotations
Train	5,365	79,335
Val	520	8,844
Test	1,197	21,006

4.2. Implementation Details

All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU (24 GB VRAM), AMD EPYC 7542 CPU, 127 GB RAM, PyTorch 2.5.1, CUDA 12.4.0, and Python 3.12. The SGD optimizer is used with an initial learning rate of 0.01, momentum of 0.937, weight decay of 0.0005, batch size of 32, and training for 300 epochs.

4.3. Evaluation Metrics

Precision (P), Recall (R), mAP@0.5 and mAP@0.5:0.95, parameter count (Params/M), and computational cost (GFLOPs) are used as evaluation metrics. mAP@0.5 represents mean AP at an IoU threshold of 50%, while mAP@0.5:0.95 averages over IoU thresholds from 50% to 95%, jointly reflecting comprehensive detection and localization performance.

4.4. Ablation Study

A stepwise cumulative ablation study is designed with YOLO11n as the baseline to verify the effectiveness and synergistic effects of MSFHA, ISCRConv, and DTGHDetect. Results are presented in Table 2.

Table 2. Ablation study results

YOLO11n	MSFHA	ISCRConv	DTGHDetect	Params/M	GFLOPs	P/%	R/%	mAP50	mAP50-95
✓				2.590	6.4	58.6	38.4	42.4	19.6
✓	✓			2.655	6.7	55.2	40.9	43.8	20.0
✓	✓	✓		2.376	6.3	57.9	41.0	44.5	20.3
✓	✓	✓	✓	2.452	6.5	58.2	41.2	45.0	20.6

Adding MSFHA to the baseline improves mAP@0.5 from 42.4% to 43.8% (+1.4%) and Recall from 38.4% to 40.9% (+2.5%), demonstrating that MSFHA effectively enhances attention to small-target edges. The slight Precision decrease (58.6%→55.2%) is attributed to the high-frequency attention activating some background regions with similar frequency characteristics to pedestrians. Adding ISCRConv further improves mAP@0.5 to 44.5% while reducing parameters to 2.376M (−8.3%) and GFLOPs to 6.3, validating the excellent performance-

ISSN: 1813-4890

efficiency trade-off of ISCRConv. The full model with DTGHDetect achieves the best performance: P=58.2%, R=41.2%, mAP@0.5=45.0%, mAP@0.5:0.95=20.6%, confirming the synergistic complementarity of all three modules.

4.5. Comparison with State-of-the-Art

Table 3 compares MID-YOLO11 with mainstream YOLO lightweight models under identical experimental protocols. Compared with YOLOv8n and YOLOv10n, the proposed model achieves mAP@0.5 improvements of +6.6% and +7.1% respectively with fewer parameters and lower GFLOPs. Compared with the latest YOLO12n, the proposed model reduces parameters by 4.5% and improves mAP@0.5 by 4.7% and Recall by 5.4%. While YOLO11s achieves a higher mAP@0.5 of 48.3% with 3.8× more parameters and 3.3× higher GFLOPs, MID-YOLO11 achieves comparable recall at a fraction of the computational cost, demonstrating a better accuracy-efficiency trade-off for resource-constrained UAV edge deployment.

Table 3. Comparison with mainstream lightweight models

Model	Params/M	GFLOPs	P/%	R/%	mAP50	mAP50-95
YOLOv8n	3.011	8.2	56.2	38.7	42.2	19.4
YOLOv10n	2.707	8.4	55.1	38.7	42.0	19.6
YOLO11n	2.590	6.4	58.6	38.4	42.4	19.6
YOLO11s	9.428	21.5	62.3	43.5	48.3	23.2
YOLO12n	2.568	6.5	57.1	39.1	43.0	19.8
MID-YOLO11 (Ours)	2.452	6.5	58.2	41.2	45.0	20.6

4.6. Visualization Analysis

To intuitively demonstrate the effectiveness of MID-YOLO11, Eigen-CAM heatmap visualization is employed to analyze the feature response of the model detection head (Fig. 5). The heatmap of the baseline YOLO11n shows a relatively diffuse distribution with high-energy responses covering background regions around targets. In contrast, MID-YOLO11 produces highly focused heatmaps with deep red regions tightly concentrated on pedestrian targets, while background regions show near-zero response. This improvement stems from the synergistic effect of MSFHA and DGTI: MSFHA assigns extremely low weights to background regions, blocking noise propagation at the feature extraction source; DGTI uses the spatial mask from the localization branch to explicitly guide the classification branch, forcing it to focus on high-confidence localization regions even for distant small targets.

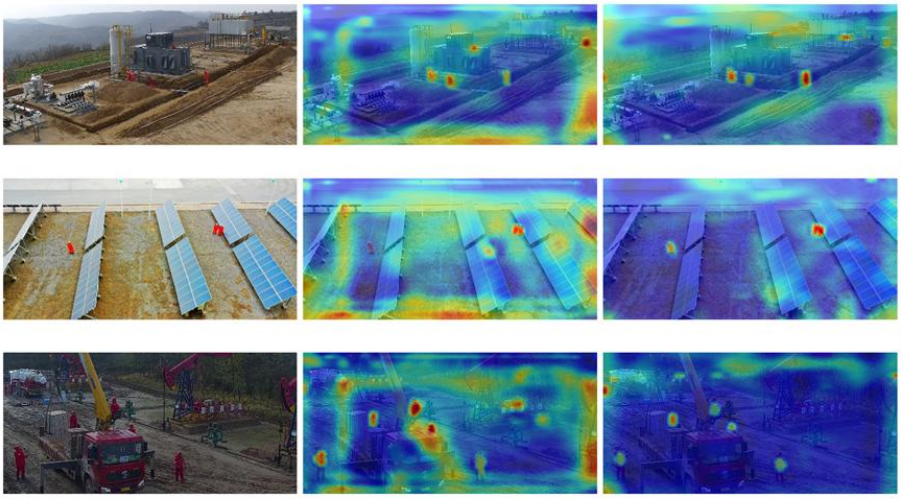

Fig. 5. Eigen-CAM heatmap comparison (Left: Original image, Middle: YOLO11n, Right: MID-YOLO11)

ISSN: 1813-4890

5. Summary

This paper proposes MID-YOLO11, a lightweight and efficient object detection model for small-target pedestrian detection in UAV aerial photography. Three core improvements — MSFHA attention mechanism, ISCRConv integrated spatial-channel reconstruction convolution, and DTGHDetect dual-gated task-guided detection head — construct a complete optimization framework for UAV aerial photography scenarios.

On the preprocessed VisDrone2019 pedestrian detection subset, MID-YOLO11 improves mAP@0.5 from 42.4% to 45.0% (+2.6 percentage points) and mAP@0.5:0.95 from 19.6% to 20.6% (+1.0 percentage point), with parameters reduced to 2.45M and GFLOPs of 6.5. Ablation experiments verify the effectiveness and synergistic gains of all three modules. Comparative experiments demonstrate MID-YOLO11 comprehensively outperforms mainstream lightweight models including YOLOv8n, YOLOv10n, and YOLO12n under comparable parameter counts. Heatmap visualization intuitively reveals the mechanism by which the improvement modules enhance feature discriminability.

In conclusion, MID-YOLO11 achieves an excellent balance between accuracy and efficiency, providing a practical lightweight detection solution for resource-constrained UAV edge deployment. Future work will further explore model generalization under more complex meteorological conditions and multi-class target scenarios.

References

[1] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[2] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[3] 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. 2023: 7464-7475.

[4] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in neural information processing systems, 2024, 37: 107984-108011.

[5] Jocher G, Chaurasia A, Qiu J. Ultralytics[J]. GitHub Repository, 2023.

[6] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2

[7] Yang L, Zhang R Y, Li L, et al. Simam: A simple, parameter-free attention module for convolutional neural networks[C]//International conference on machine learning. PMLR, 2021: 11863-11874.

[8] Li J, Wen Y, He L. Scconv: Spatial and channel reconstruction convolution for feature redundancy[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 6153-6162.

[9] 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.