ISSN: 2710-0588

MSG-DETR: A Small Object Detection Algorithm for UAV Aerial Images

Xiang Li, Ruxin Gao*

School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo, 454000, China

*Corresponding author:(Email: gaoruxin@hpu.edu.cn)

Abstract

Due to the significant scale variations and dense distribution of small objects commonly present in imagery captured by unmanned aerial vehicles (UAV), traditional object detection algorithms often suffer from missed detections and false positives in such scenarios. To address these challenges, we propose a novel detection framework—MSG-DETR, specifically designed for small object detection in aerial images captured by UAV. First, we design a lightweight multi-scale feature fusion backbone, MSFFNet, which enhances the extraction of small object features while significantly reducing computational overhead. Second, we introduce the Small-object Feature Fusion (SF-Fusion), which incorporates rich P3-level features from MSFFNet into the neck architecture to deepen feature fusion and mitigate information loss. Finally, we integrate a Gated Convolutional Attention Mechanism (GCAM) to improve the model's ability to perceive and localize tiny objects in cluttered backgrounds. Experimental results on the VisDrone2019 dataset demonstrate that MSG-DETR achieves performance gains of +3.8% in mAP@0.5 and +3.5% in mAP@0.5:0.95, while reducing the number of parameters by 28.1% compared to the baseline model.

Keywords

UAV object detection; small object detection; feature fusion; attention mechanism.

1. Introduction

With the rapid advancement and widespread deployment of UAV technologies, UAV have been extensively applied in various domains, including urban traffic monitoring [1], military reconnaissance [2], and aerial surveying [3]. However, UAV-captured images often exhibit significant scale variations and densely packed small objects, posing substantial challenges to object detection algorithms [4]. For instance, in aerial views, small and densely distributed objects such as bicycles and pedestrians often occupy only a few pixels and are easily missed by conventional detectors. Moreover, the onboard computational resources and energy consumption of UAV platforms are highly constrained, necessitating the design of lightweight and efficient detection models.

Traditional object detection methods struggle to balance detection accuracy and inference speed in UAV-based scenarios. Two-stage detectors, such as Faster R-CNN [5], typically achieve high accuracy through region proposal mechanisms but suffer from slow inference and high computational overhead, making them unsuitable for real-time UAV applications. In contrast, one-stage detectors like YOLO [6] and SSD [7] offer faster inference and simpler architectures but often lack the fine-grained feature extraction necessary for detecting small objects in UAV imagery, leading to suboptimal performance. Consequently, both paradigms face inherent trade-offs between speed and accuracy in UAV object detection tasks. Compared to traditional CNN-based detectors, end-to-end Transformer-based frameworks (e.g., DETR [8]) eliminate

ISSN: 2710-0588

hand-crafted components by leveraging self-attention mechanisms, enabling stronger global modeling capabilities. However, the original DETR suffers from slow convergence and high computational cost. To address these limitations, Zhao et al. [9] proposed RT-DETR, a real-time variant that achieves superior performance over mainstream YOLO models in both accuracy and speed. Despite its efficiency improvements, RT-DETR still presents several limitations: it employs a ResNet-based single-path backbone that lacks adaptive modeling of multi-scale objects; its simplified bidirectional feature fusion structure, while computationally efficient, weakens the ability to capture long-range dependencies, resulting in performance degradation in complex scenes; traditional convolution operations in the neck stage struggle to integrate global contextual information effectively, making the model prone to missed or false detections when dealing with occluded or densely packed small objects.

To overcome these issues, we propose MSG-DETR, a small object detection framework tailored for UAV-captured imagery. The proposed method achieves enhanced detection accuracy while maintaining low computational and storage costs. The main contributions of this work are summarized as follows:

A lightweight Multi-Scale Feature Fusion Network (MSFFNet) is proposed to effectively reduce computational redundancy while enhancing feature extraction. This design significantly improves the capability of extracting small object features and greatly reduces computational overhead, making it suitable for UAV-based detection tasks.

A Small-object Feature Fusion module (SF-Fusion) is developed to address the challenge of increased computation and parameter size caused by adding low-level detection layers such as P2 in conventional methods. By effectively integrating small-object features from the backbone into the neck structure, this module deepens the fusion process and minimizes information loss, leading to improved detection performance in UAV imagery.

A Gated Convolutional Attention Mechanism (GCAM) is introduced to enhance the model's sensitivity to small objects in complex backgrounds. By providing stronger global modeling capabilities, GCAM improves the perception and localization of small targets, thereby boosting detection accuracy in dense and cluttered scenes.

2. Related Work

2.1. Traditional Object Detection Methods

Early object detectors primarily relied on manually designed anchor boxes and post-processing operations. Two-stage approaches, such as Faster R-CNN, generate region proposals and classify them to achieve high detection accuracy. However, their multi-step pipeline limits their real-time deployment capabilities. In contrast, one-stage detectors like the YOLO series and SSD directly regress object locations and categories in a single forward pass, offering faster inference but often at the cost of reduced accuracy. Regardless of the paradigm, the use of hand-crafted components such as anchor boxes and non-maximum suppression (NMS) can become bottlenecks, particularly in scenarios involving multi-scale and densely packed objects. Anchor-free detectors such as TOOD [10] aim to simplify network design by regressing the four boundaries of objects directly, yet they still struggle to detect small objects effectively. Overall, traditional CNN-based methods show limited capability in capturing tiny targets within UAV imagery.

2.2. Transformer-Based Object Detection

Transformer-based detectors, particularly the DETR family, eliminate many hand-crafted components through self-attention mechanisms, offering a novel end-to-end object detection paradigm. However, the original DETR suffers from slow convergence, high computational cost, and insufficient support for multi-scale detection. To address these limitations, Deformable

ISSN: 2710-0588

DETR [11] introduced sparse attention mechanisms that focus only on salient regions of the feature maps, significantly accelerating training convergence. RT-DETR further improved detection performance and inference speed by incorporating an efficient backbone and a streamlined decoder structure, achieving real-time performance comparable to the YOLO series. Despite these advances, RT-DETR still requires considerable computational resources and has a relatively large parameter count, which hinders its deployment on edge devices or mobile platforms. This computational burden remains a key bottleneck for its application in real-world UAV-based object detection systems.

2.3. Small Object Detection

To address the challenges of small object detection, numerous methods have been proposed to enhance detection accuracy and robustness. Lin et al. [12] introduced the Feature Pyramid Network (FPN), which employs a top-down architecture to enrich high-resolution semantic features, significantly improving small object detection performance. Tan et al. [13] proposed EfficientDet, which utilizes a bidirectional feature pyramid network (BiFPN) for efficient multi-scale fusion, achieving a good balance between accuracy and speed. SOD-YOLO [14] designed a novel module that fuses shallow and deep features while avoiding excessive use of group and pointwise convolutions, reducing memory access overhead and improving feature richness for small objects. LEAF-YOLO [15] enhanced small object representation by incorporating high-resolution shallow features and optimizing fusion strategies. TPH-YOLOv5 [16] integrated the Transformer architecture and CBAM attention into YOLOv5, improving detection performance in dense target scenarios. ETAM [17] introduced a magnifying glass structure and a quadruple attention mechanism to boost the expressiveness of small object features. AO2-DETR, proposed by Dai et al. [18], employed an oriented proposal generation mechanism to enhance decoder interaction with spatial features, enabling direction-aware detection. Ren et al. [19] combined a super-resolution generation module with adversarial learning to improve detection performance under low-resolution and complex background conditions. Wang et al. [20] presented AMFEF-DETR, which uses frequency-adaptive dilated convolution for dynamic feature extraction and HiLo attention to strengthen interactions between high- and low-frequency components. Zhang et al. [21] designed a detection module based on multi-scale atrous convolution, which expands the receptive field without increasing parameters and enhances the modeling of scale and spatial variations, significantly improving small object detection accuracy and robustness in UAV imagery.

3. Method

3.1. Overall Architecture of MSG-DETR

MSG-DETR is proposed as a novel framework for small object detection, aiming to improve accuracy and robustness under complex interference scenarios from a UAV perspective. As shown in Fig. 1, the overall architecture comprises the MSFFNet backbone, encoder, and decoder working collaboratively. By effectively exploiting multi-scale features and enhancing feature representation capability, the framework significantly boosts small object detection performance and provides an efficient and reliable solution for challenging UAV-based detection tasks.

ISSN: 2710-0588

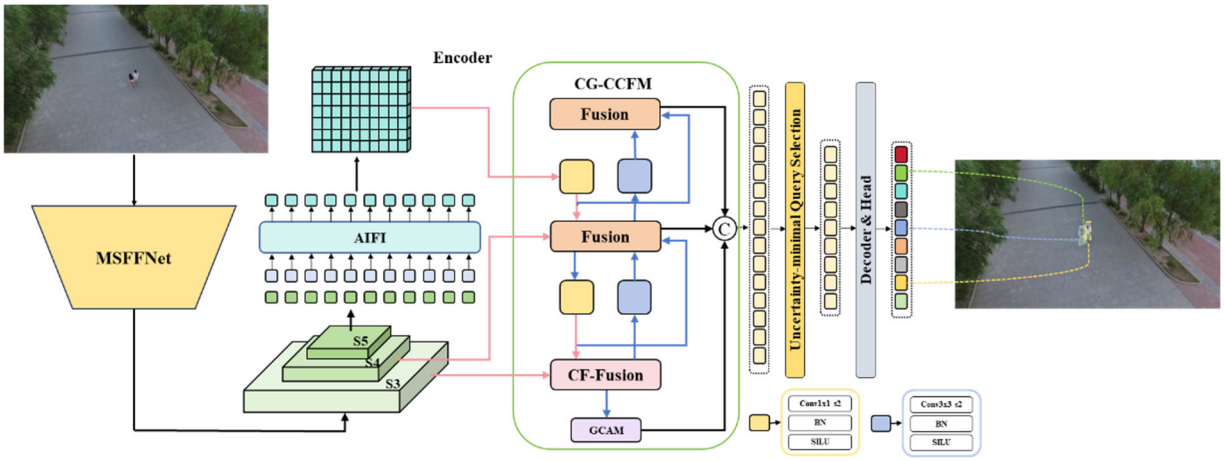

Figure 1. MSG-DETR Architecture Diagram

3.2. MSFFNet

RT-DETR employs ResNet as its backbone network; however, several limitations emerge in UAV-based object detection scenarios. Specifically, the limited local receptive field of ResNet hampers the extraction of fine-grained features from small objects. In addition, high-frequency details tend to be lost in deeper layers, degrading localization accuracy in complex environments. Furthermore, the large number of parameters in ResNet poses challenges for deployment on resource-constrained platforms. To address these issues, a multi-scale feature fusion backbone network (MSFFNet) is proposed to enhance feature representation while maintaining a lightweight design. As illustrated in Fig. 2, MSFFNet consists of two main components: the Multi-Scale Feature Fusion (MSFF) module and a standard convolution (Conv) module.

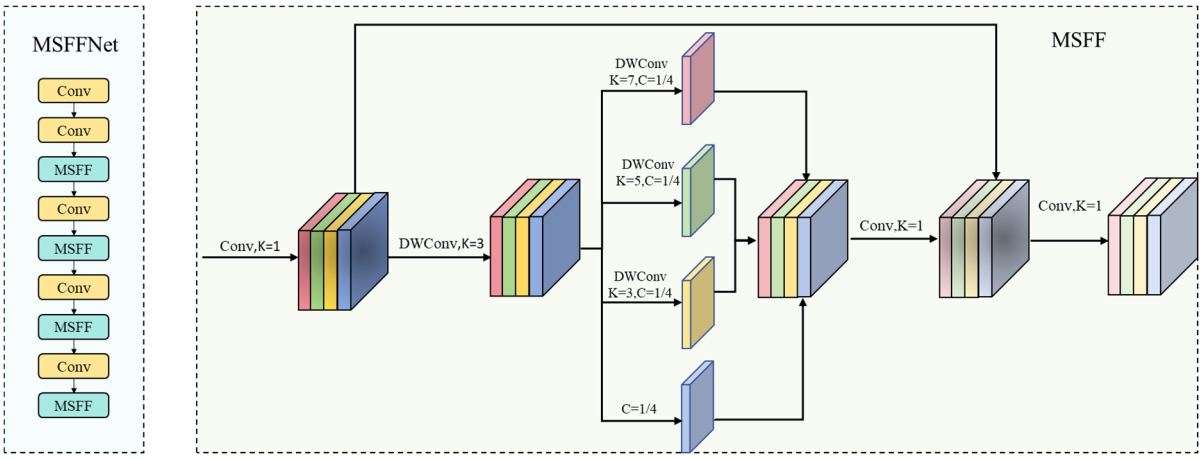

Figure 2. MSFFNet and MSFF Architecture Diagram

The MSFF module adopts a dual-branch design to simultaneously preserve original information and enhance multi-scale feature extraction. One branch directly forwards the input features to the output to retain low-level spatial information. The other branch begins with a 3×3 depthwise separable convolution(DWConv)[22] to extract efficient local features, followed by a channel-wise division of the feature map into four equal groups. Each group is then processed in parallel using depthwise separable convolutions with different kernel sizes (3×3, 5×5, and 7×7) to capture information across various receptive fields. The resulting multi-scale features are concatenated along the channel dimension and further fused using a 1×1 convolution to enhance contextual representation. Finally, features from both branches are concatenated and

ISSN: 2710-0588

integrated, enabling effective utilization of multi-scale information for robust feature representation.

The MSFF module adopts a multi-branch architecture, which processes multi-scale feature streams in parallel to significantly enhance the extraction of scale-aware features. Compared to conventional single-branch designs, this structure demonstrates superior performance in small object detection tasks by effectively capturing features across varying spatial resolutions and improving both detection accuracy and robustness.

3.3. SF-Fusion

In the RT-DETR architecture, the Fusion component within the CCFM module integrates multi-scale feature maps to facilitate cross-level information aggregation. This process features a simple design and high computational efficiency, making it well-suited for deployment in resource-constrained environments. However, its capability to model fine-grained details is limited, which restricts performance under complex scenes or in small object detection tasks. To enhance detection accuracy for small objects, existing approaches often incorporate an additional P2 detection head to leverage high-resolution features. Nevertheless, this strategy significantly increases computational cost and inference latency. To address these challenges, an improved feature fusion mechanism, termed SF-Fusion, is proposed based on the original Fusion structure, aiming to strengthen the model's sensitivity to small objects.As illustrated in Fig. 3.

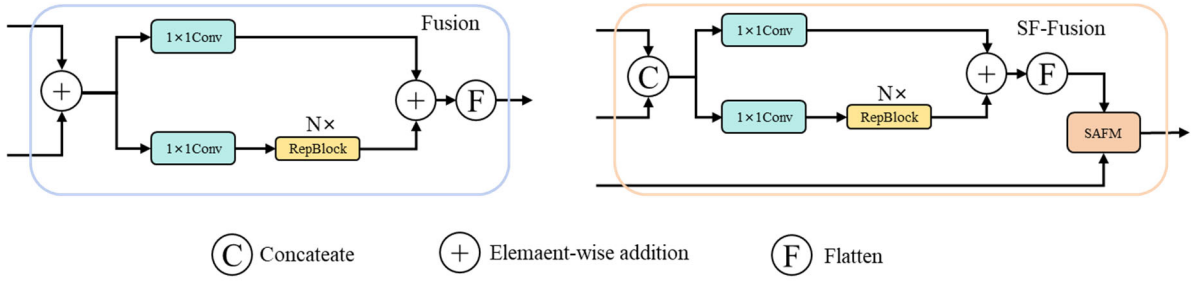

Figure 3. Fusion and SF-Fusion Architecture Diagram

a Small-object Aware Feature Fusion Module (SAFM) is introduced to integrate shallow feature maps—rich in small-object information—from the backbone into the CCFM module. This design effectively leverages fine-grained features from earlier layers and fuses them with deeper semantic representations, thereby enabling more expressive multi-scale feature representations. The detailed structure of the SAFM module is shown in Fig. 4.

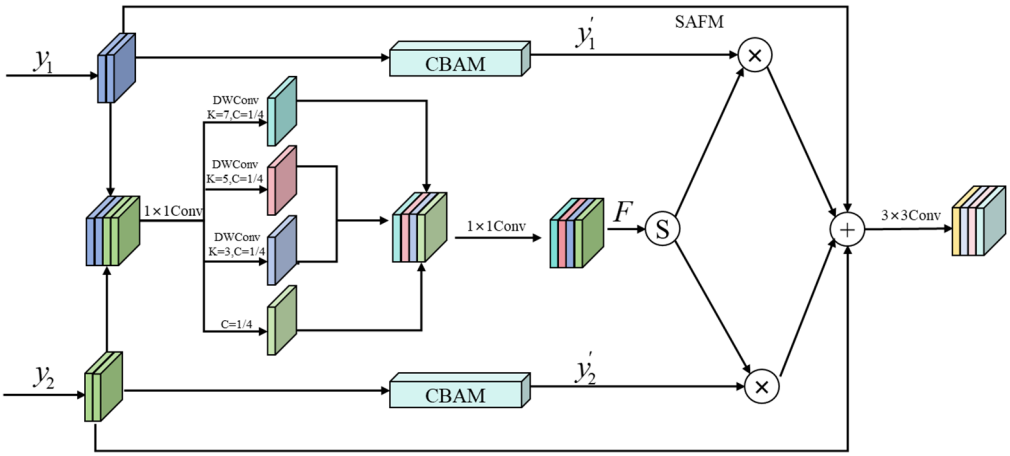

Figure 4. SAFM Architecture Diagram

ISSN: 2710-0588

Specifically, the SAFM module takes two feature maps with different semantic levels as input. These features are first concatenated along the channel dimension and passed through a 1×1 convolution to reduce dimensionality and obtain an initial fused representation. The resulting features are then split into four channel-wise branches. Among them, three branches employ depthwise separable convolutions with kernel sizes of 3×3, 5×5, and 7×7 respectively, to capture contextual information under varying receptive fields. The fourth branch preserves the original features to enhance residual representation capability. The outputs from all four branches are concatenated and further integrated using another 1×1 convolution, resulting in the fused feature map F:

$$F = \mathrm{Conv}_{1\times1}\left(\mathrm{Concat}\left(\mathrm{DWConv}_{3\times3}(x_1), \mathrm{DWConv}_{5\times5}(x_2), \mathrm{DWConv}_{7\times7}(x_3), x_4\right)\right) \quad (1)$$

The fusion weight $A = \sigma F$ is obtained through the Sigmoid activation function to adjust the importance of different input branches. Prior to feature fusion, in order to enhance the network's perception of key local regions, the CBAM [23] attention mechanism is applied to the input features y 1 and $y2$, respectively.

SF-Fusion mitigates feature loss by integrating small object features from different layers, thereby enhancing feature representation and improving detection accuracy.

3.4. GCAM

The neck of RT-DETR plays a crucial role in feature fusion and multi-scale representation. The P3 layer typically corresponds to shallower, higher-resolution feature maps and is primarily responsible for detecting small objects. However, due to its lower semantic level and the presence of significant background noise, traditional convolutional operations often struggle to effectively model contextual relationships between objects at this scale. This limitation becomes more pronounced in complex environments with multi-scale objects and background interference, leading to redundant information and unstable feature responses, which ultimately degrade detection performance. To address this issue, a Gated Convolutional Attention Module (GCAM) is introduced for the P3 layer in the neck of RT-DETR, as shown in Fig. 5.

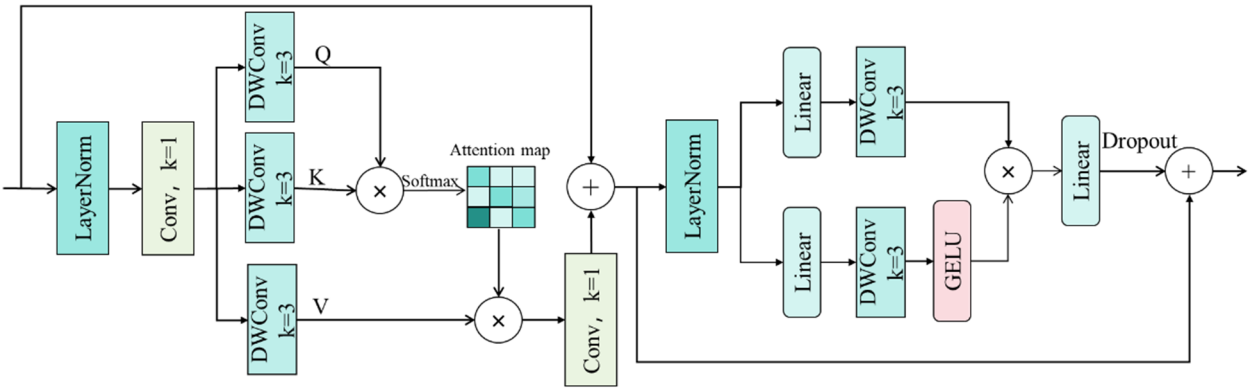

Figure 5. GCAMArchitecture Diagram

The input feature map *X* is sequentially processed through a convolutional attention module followed by a gated convolutional structure. Convolutional Attention Path: First, the input is normalized using LayerNorm. Then, a 1×1 convolution is applied to generate three sets of features: query (Q), key (K), and value (V). To enhance local perceptual capability, depthwise convolution (DWConv) is incorporated. A multi-head attention mechanism is then used to compute attention weights, guiding the network to focus on key regions within the image and

ISSN: 2710-0588

improving the selective representation of spatial features. Finally, the output is projected back to the original channel dimension using a 1×1 convolution and fused with the input via a residual connection to stabilize training and preserve the original feature information.

Gated Convolutional Linear Module (GCLM): In this path, the input features are first normalized using LayerNorm, followed by a linear projection that expands the channel dimension to twice its original size. The resulting feature map is then split into two branches, each processed by a depthwise separable convolution to model spatial relationships. The outputs of the two branches are fused via element-wise multiplication, enabling joint modeling of spatial and channel-wise dependencies. The fused representation is subsequently projected back to the original channel dimension through a linear layer and regularized with a Dropout operation to pr This structure captures both local spatial details and global contextual dependencies, thereby enhancing the model's capacity to represent complex scenes. In this study, the GCAM module is embedded into the P3 layer of the neck, resulting in a significant improvement in small object detection accuracy and robustness under challenging environments.event overfitting and enhance generalization.

4. Experiments and Results Analysis

4.1. Dataset

The dataset used in this paper is the public UAV dataset VisDrone2019[24].

4.2. Experimental Setup

All experiments were conducted on a Windows 10 workstation equipped with an Intel i7-12700F CPU and an NVIDIA GeForce RTX 3080 GPU. The PyTorch 1.12.0 framework was used for model training and evaluation.

4.3. Ablation Study

To evaluate the impact of the proposed modules on model performance, ablation studies were conducted on the VisDrone2019 dataset focusing on the identified improvements. The results are summarized in Table 1, where A, B, and C denote the MSFFNet, SF-Fusion, and GCAM modules, respectively, and the symbol "✓" indicates the inclusion of the corresponding module.

Table 1. Ablation Study on the VisDrone2019 Dataset

A	B	C	mAP0.5%	mAP0.5:0.95%	Parameters/M	GFLOPs	Weight/MB	FPS
			46.0	27.2	19.9	57.0	40.5	83.3
√			47.9	29.2	12.8	42.9	25.2	84.6
	√		48.5	29.9	20.7	67.6	40.2	78.4
		√	47.7	29.2	20.6	65.6	39.9	76.9
√	√		48.9	30.2	13.6	53.1	26.8	78.3
√		√	47.7	29.0	13.5	51.5	26.5	77.3
	√	√	48.8	30.1	21.4	76.2	41.5	68.5
√	√	√	49.8	30.7	14.3	61.6	28.0	71.7

After incorporating module A, mAP@0.5 increased to 47.9% (+1.9%) and mAP@0.5:0.95 rose to 29.2% (+2.0%), while the number of parameters decreased by 35.6% (to 12.8M) and FLOPs were reduced by 24.7% (to 42.9 GFLOPs), demonstrating the dual benefits of enhanced small object feature extraction and model lightweight design. When modules B and C were used individually, mAP@0.5 improved to 48.5% (+2.5%) and 47.7% (+1.7%) respectively, albeit with a slight increase in parameters and computation, indicating the need for lightweight design optimizations to manage resource consumption. The combined use of modules A and B

ISSN: 2710-0588

yielded a significant mAP@0.5 improvement to 48.9% (+2.9%), with parameters and FLOPs controlled at 13.6M and 53.1 GFLOPs, highlighting their complementary effects in balancing lightweight design and information preservation. Integrating all modules resulted in optimal performance, achieving 49.8% mAP@0.5 and 30.7% mAP@0.5:0.95, with 14.3M parameters, 67.0 GFLOPs computation cost, and a model size of only 28.0MB, demonstrating a balanced optimization of accuracy and efficiency through module collaboration.

4.4. Comparative Experiments

In the comparative experiments conducted on the VisDrone2019 dataset, as shown in Table 2, MSG-DETR demonstrates significant advantages over several state-of-the-art models. Compared to RT-DETR-R18, MSG-DETR achieves 49.8% mAP@0.5 and 31.4% mAP@0.5:0.95 on the validation set, surpassing RT-DETR-R18's 46.0% and 27.7%, respectively. On the test set, MSG-DETR attains 39.6% mAP@0.5 and 23.2% mAP@0.5:0.95. Additionally, YOLOv12X [25], with 59.1M parameters, only achieves 46.4% mAP@0.5 and 28.8% mAP@0.5:0.95, making MSG-DETR superior by 3.4% and 2.6%, respectively, while maintaining a more lightweight design. Moreover, MSG-DETR outperforms both single-stage detectors (e.g., TOOD, GFI [26], and DINO [27]) and two-stage frameworks (e.g., Cascade R-CNN [28] and Faster R-CNN) in terms of both accuracy and inference efficiency, indicating its suitability and competitiveness for UAV-based object detection tasks.

Table 2. Comparative Experiments on the VisDrone2019 Dataset

Model	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5-0.95/%	Parameters/M	Weight/MB
YOLOv12-X	46.4/38.1	28.8/22.9	59.1	198.6	119.1	58.01
TOOD	39.2/33.3	23.9/19.6	32.0	199.0	-	40.3
GFL	37.1/22.6	31.7/18.4	32.3	206	-	33.7
DINO	49.5/40.8	28.5/22.6	47.6	274.0	-	15.1
Faster R-CNN	12.5/10.1	6.6/5.2	41.4	208	-	32.5
Cascade r-cnn	28.3/32.1	23.3/18.9	62.3	236	-	31.6
RT-DETR-R18	46.0/37.2	27.7/21.1	19.9	57.0	40.5	83.3
RT-DETR-R34	47.9/38.4	29.4/22.3	31.1	88.8	63.0	73.0
MSG-DETR	49.8/39.6	30.7/23.2	14.3	67.0	28.0	71.7

4.5. Visualization Results

On the test set of the VisDrone2019 dataset, visual inference analysis of MSG-DETR is conducted in this paper, and the results are shown in Figure 6. Among them, green boxes represent correct detections, blue boxes denote false detections, and red boxes indicate missed detections. It can be observed from the results that MSG-DETR achieves a significant improvement in detection accuracy when handling scenarios with multi-scale and dense targets. Meanwhile, the number of false detections and missed detections is obviously reduced under complex background environments. This demonstrates that the proposed model possesses stronger adaptability and robustness when coping with challenges such as occlusion and background interference, which verifies its detection efficiency and reliability in practical complex scenarios.

Figure 6. Detection results of guide rail inspection video

ISSN: 2710-0588

5. Conclusion

This paper innovatively proposes an end-to-end object detection model named MSG-DETR for UAV aerial images. The model integrates a lightweight multi-scale feature fusion backbone network (MSFFNet), a small object feature fusion module (SF-Fusion), and a gated convolutional attention mechanism (GCAM). It achieves superior detection performance on the representative VisDrone2019 dataset, with mAP50 reaching 49.8% and mAP@0.5:0.95 attaining 30.7%.

Compared with the baseline model RT-DETR-R18, MSG-DETR reduces the parameter count by 22%, while increasing mAP@0.5 and mAP@0.5:0.95 by 3.8% and 3.5%, respectively. It balances detection accuracy and efficiency while maintaining real-time inference speed, demonstrating great potential and practical application value in resource-constrained scenarios.

References

[1] Feng J, Wang J, Qin R. Lightweight detection network for arbitrary-oriented vehicles in UAV imagery via precise positional information encoding and bidirectional feature fusion[J]. International Journal of Remote Sensing, 2023, 44(15): 4529-4558.

[2] Qu Y, Sun H, Dong C, et al. Elastic collaborative edge intelligence for UAV swarm: Architecture, challenges, and opportunities[J]. IEEE Communications Magazine, 2023, 62(1): 62-68.

[3] Wang X, Demartino C, Narazaki Y, et al. Rapid seismic risk assessment of bridges using UAV aerial photogrammetry[J]. Engineering Structures, 2023, 279: 115589.

[4] Du B, Huang Y, Chen J, et al. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 13435-13444.

[5] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149.

[6] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[7] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Cham: Springer International Publishing, 2016: 21-37.

[8] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229.

[9] Zhao Y, Lv W, Xu S, et al. Detrs beat yolos on real-time object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 16965-16974.

[10] Feng C, Zhong Y, Gao Y, et al. Tood: Task-aligned one-stage object detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society, 2021: 3490-3499.

[11] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection[J]. arXiv preprint arXiv:2010.04159, 2020.

[12] 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: 2117-2125.

[13] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.

[14] Xiao Y, Di N. SOD-YOLO: A lightweight small object detection framework[J]. Scientific Reports, 2024, 14(1): 25624.

[15] Nguyen H H, Hoang M S. LEAF-YOLO: Lightweight Edge-Real-Time Small Object Detection on Aerial Imagery[J]. Intelligent Systems with Applications, 2025, 25: 200484.

ISSN: 2710-0588

[16] Yin Y, Yu J, Chen P, et al. Road crack detection of drone-captured images based on TPH-YOLOv5[J]. International Journal of Pavement Engineering, 2025, 26(1): 2474729.

[17] Zhang J, Xia K, Huang Z, et al. ETAM: Ensemble transformer with attention modules for detection of small objects[J]. Expert systems with applications, 2023, 224: 119997.

[18] Dai L, Liu H, Tang H, et al. AO2-DETR: Arbitrary-oriented object detection transformer[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(5): 2342-2356.

[19] Ren K, Gao Y, Wan M, et al. Infrared small target detection via region super resolution generative adversarial network[J]. Applied Intelligence, 2022, 52(10): 11725-11737.

[20] Wang S, Jiang H, Yang J, et al. Amfef-detr: An end-to-end adaptive multi-scale feature extraction and fusion object detection network based on uav aerial images[J]. Drones, 2024, 8(10): 523.

[21] Zhang Y, Jia R S, Yang R, et al. DSNet: A vehicle density estimation network based on multi-scale sensing of vehicle density in video images[J]. Expert Systems with Applications, 2023, 234: 121020.

[22] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.

[23] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[24] 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. 2019: 0-0.

[25] Tian Y, Ye Q, Doermann D. Yolov12: Attention-centric real-time object detectors[J]. arXiv preprint arXiv:2502.12524, 2025.

[26] Hu Y, Zhou Y, Xiao J, et al. GFL: A decentralized federated learning framework based on blockchain[J]. arXiv preprint arXiv:2010.10996, 2020.

[27] Zhang H, Li F, Liu S, et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection[J]. arXiv preprint arXiv:2203.03605, 2022.

[28] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6154-6162.