frontiers | Frontiers in Remote Sensing

TYPE Original Research
PUBLISHED 22 July 2025
DOI 10.3389/frsen.2025.1599099

OPEN ACCESS

EDITED BY
Xinghua Li,
Wuhan University, China

REVIEWED BY
Krzysztof Karsznia,
Warsaw University of Technology, Poland
Tao Ning,
Dalian Nationalities University, China
Xuesong Jiang,
Qilu University of Technology, China

*CORRESPONDENCE
Zhizhao Zhang,
472321772@stu.lntu.edu.cn

RECEIVED 24 March 2025
ACCEPTED 03 July 2025
PUBLISHED 22 July 2025

CITATION
Zhu X and Zhang Z (2025) Efficient vision transformers with edge enhancement for robust small target detection in drone-based remote sensing.
Front. Remote Sens. 6:1599099.
doi: 10.3389/frsen.2025.1599099

Efficient vision transformers with edge enhancement for robust small target detection in drone-based remote sensing

Xuguang Zhu[1] and Zhizhao Zhang[2]*

[1]College of Innovation and Practice, Liaoning Technical University, Fuxin, China, [2]School of Software, Liaoning Technical University, Huludao, China

Small object detection in UAV remote sensing imagery faces significant challenges due to scale variations, background clutter, and real-time processing requirements. This study proposes a lightweight transformer-based detector, MLD-DETR, which enhances detection performance in complex scenarios through multi-scale edge enhancement and hierarchical attention mechanisms. First, a Multi-Scale Edge Enhancement Fusion (MSEEF) module is designed, integrating adaptive pooling and edge-aware convolution to preserve target boundary details while enabling cross-scale feature interaction. Second, a Layered Attention Fusion (LAF) mechanism is developed, leveraging spatial depth-wise convolution and omnidirectional kernel feature fusion to improve hierarchical localization capability for densely occluded targets. Furthermore, a Dynamic Positional Encoding (DPE) module replaces traditional fixed positional embeddings, enhancing spatial perception accuracy under complex geometric perspectives through learnable spatial adapters. Combined with an Inner Generalized Intersection-over-Union (Inner-GIoU) loss function to optimize bounding box geometric consistency, MLD-DETR achieves 36.7% AP50% and 14.5% APs on the VisDrone2019 dataset, outperforming the baseline RT-DETR by 3.2% and 1.8% in accuracy while achieving 20% parameter reduction and maintaining computational efficiency suitable for UAV platforms equipped with modern edge computing hardware. Experimental results demonstrate the algorithm's superior performance in UAV remote sensing applications such as crop disease monitoring and traffic congestion detection, offering an efficient solution for real-time edge-device deployment.

KEYWORDS

UAV, drone-based remote sensing, RT-DETR, small object detection, multi-scale edge enhancement

1 Introduction

The rapid advancement of drone-based aerial imaging has brought about significant transformations across various industries. With improvements in drone technology, an increasing number of sectors are utilizing drones for cost-effective and efficient aerial data collection, including agricultural monitoring (Zhang et al., 2021), urban planning, disaster assessment, traffic surveillance, and environmental protection. Drones, equipped with high-resolution cameras, are capable of capturing expansive images and providing valuable visual data from hard-to-reach or complex areas, which proves essential for a wide range of applications. Traditional object detection frameworks like Faster R-CNN (Ren et al., 2016)

and YOLO (Redmon et al., 2016) rely heavily on anchor-based proposals and Non-Maximum Suppression (NMS) for duplicate removal. While effective in general scenarios, these components become limiting factors in drone imagery: predefined anchors struggle with extreme scale variations (e.g., 0.5 m–500 m altitudes), and NMS frequently fails in dense object clusters common to aerial views.

RT-DETR (Zhao et al., 2024) addresses these issues through an end-to-end transformer architecture that eliminates anchors and NMS via learnable queries. Moreover, its hybrid encoder—fusing CNN's robust local feature extraction with the transformer's global reasoning—significantly boosts detection performance on challenging aerial imagery. However, when it comes to small target detection in aerial imagery, several issues still need to be addressed:

1. Limited Small Object Representation: RT-DETR excels in global feature aggregation but lacks mechanisms tailored to preserving the fine details of small objects. The inevitable downsampling in CNN layers may lead to the loss of crucial edge and texture information, thereby reducing detection accuracy for small targets;
2. Inadequate Multi-Scale Feature Fusion: While incorporating CNN-based local feature extraction, RT-DETR's strategy for fusing features across multiple scales is not fully optimized for aerial scenarios. Small targets captured at varying altitudes and perspectives require more robust multi-scale integration to reliably capture their details;
3. Sparse Query Mechanism in Dense Scenes: The learnable query mechanism, effective in many contexts, may not adequately cover densely packed small targets. This sparsity in query allocation can result in missed detections when multiple small objects are clustered closely together;
4. Suboptimal Positional Encoding: RT-DETR typically employs standard sine-cosine positional encodings. However, the complex geometric variations inherent in drone imagery might require more adaptive or learnable encoding schemes to accurately localize small objects under varied perspectives;
5. Insufficient Edge and Context Enhancement: In aerial views, small objects often blend into busy backgrounds. Without explicit modules to enhance edge and contextual information, RT-DETR may struggle to distinguish these targets from their surroundings, especially in low contrast or occluded settings;
6. High Computational Demand Affecting Real-Time Performance: Although the removal of anchors and NMS reduces some computational overhead, the transformer-based architecture of RT-DETR still entails significant resource consumption. This high computational demand can be a bottleneck for real-time processing in drone applications.

To tackle the challenges described above, we propose MLD-DETR—a novel detection framework designed specifically for drone-based aerial imagery. MLD-DETR is built upon four key modules that work synergistically to enhance small object detection by preserving fine details, optimizing feature fusion, and refining localization. The key innovations of this model are as follows:

1. Multi-Scale Edge-Enhanced Feature Fusion (MSEEF) Module for Enhanced Feature Extraction: This module integrates multi-scale feature extraction with edge enhancement, enabling the model to capture fine-grained details at various scales while preserving object boundaries. This is especially beneficial for detecting small and partially occluded objects in complex, cluttered environments;
2. Layered Attention Fusion (LAF) Module for Optimized Small Object Detection: To overcome the limitations of traditional feature pyramids, the LAF module refines the feature hierarchy using SPDConv for small-scale feature integration combined with a CSP-OmniKernel fusion process. This design effectively enhances small object detection in high-resolution aerial imagery while maintaining computational efficiency;
3. Dynamic Position Encoding (DPE) Module for Improved Spatial Representation: Recognizing that fixed positional encodings are insufficient for the complex spatial relationships in drone imagery, the DPE module introduces a learned, adaptive positional encoding scheme. This enables the model to dynamically adjust position representations, significantly improving localization accuracy;
4. Inner-GIoU Module for Refined Localization: Traditional IoU-based loss functions may not yield precise localization, especially for small or occluded objects. The Inner-GIoU module refines the intersection-over-union computation within the inner regions of bounding boxes, enhancing geometric consistency and robustness in localization.

By integrating these four modules, MLD-DETR achieves a balanced trade-off between detection accuracy, computational efficiency, and robustness, making it specifically suitable for UAV-based object detection applications where both precision and real-time performance are critical. Unlike existing methods that address UAV detection challenges separately, our integrated approach synergistically combines multi-scale processing, edge enhancement, and adaptive positioning to tackle the unique challenges of aerial imagery in a unified framework.

The rest of the paper is organized as follows. Section 2 reviews the related literature and highlights the limitations of current approaches. In Section 3, we detail our proposed MLD-DETR model and its novel modules, including MSEEF, LAF, DPE, and the Inner-GIoU loss function. Section 4 presents extensive experimental evaluations, ablation studies, and comparative analyses, while Section 5 concludes the paper and discusses potential future research directions.

2 Related work

In the early evolution of object detection, traditional methods laid the groundwork for modern techniques. Region-based approaches such as R-CNN (Girshick et al., 2014), Fast R-CNN (Girshick et al., 2015), and Faster R-CNN (Ren et al., 2016) integrated convolutional neural networks (CNNs) with region proposal mechanisms, achieving significant improvements in detection accuracy through end-to-end training and refined feature extraction. Concurrently, single-stage detectors like YOLO (Redmon et al., 2016) and SSD (Liu et al., 2016) emerged,

emphasizing speed and efficiency by directly predicting object bounding boxes and class probabilities in one pass. Despite their remarkable performance, these traditional methods rely heavily on predefined anchor boxes and post-processing steps such as Non-Maximum Suppression (NMS), which can become bottlenecks when handling objects at multiple scales, particularly in complex aerial imagery.

The introduction of Detection Transformer (DETR) (Carion et al., 2020) marked a paradigm shift in object detection by framing it as a set prediction problem and leveraging a Transformer architecture for end-to-end detection. Despite its elegant formulation and elimination of traditional components such as anchors and NMS, the original DETR suffered from slow convergence and struggled with multi-scale and small object detection. To overcome the limitations of the original DETR, numerous variants have been proposed in recent years. For example, Deformable DETR (Zhu et al., 2021) introduced a deformable attention module that selectively samples key points, accelerating convergence and enhancing multi-scale feature capture. Building on this, Conditional DETR (Meng et al., 2021) improved the query design by incorporating conditional embeddings that adaptively leverage input features in complex scenes. In addition, variants like DAB-DETR (Qi et al., 2021) and Anchor DETR (Xu et al., 2022) further refined query embeddings and anchor designs to boost localization precision. More recently, methods such as D-FINE (Peng et al., 2024) and DEIM (Huang et al., 2024) have been proposed to address challenges in bounding box regression and matching, thereby further accelerating convergence and improving detection accuracy. Recent work such as AUHF-DETR (Guo et al., 2025) focuses on spatial attention mechanisms and wavelet convolution for UAV detection. However, our approach differs significantly by integrating multi-scale edge enhancement with dynamic positional encoding, providing a more comprehensive solution for the varied geometric perspectives and scale challenges inherent in aerial imagery.

3 Materials and methods

In this section, we describe the methodology behind our proposed MLD-DETR model for enhancing small object detection in drone imagery. We begin by reviewing the baseline RT-DETR architecture, followed by a detailed introduction of our novel modifications. Specifically, we present the MSEEF module, the LAF module, the DPE module, and the Inner-GIoU loss function. Together, these components are designed to improve feature extraction, localization precision, and overall detection performance.

3.1 RT-DETR

RT-DETR (Real-Time Detection Transformer) is an advanced object detection model designed to tackle the challenges of real-time applications. It combines the power of transformers with efficient design principles to deliver high-speed, high-accuracy object detection. The architecture integrates a backbone network, typically a Convolutional Neural Network (CNN) such as ResNet, to extract feature maps from input images. These feature maps are then processed through a transformer encoder-decoder architecture, which learns the spatial relationships and context within the image. The encoder generates key-value pairs that capture relevant features, while the decoder uses these pairs to predict object positions, class labels, and bounding boxes. Unlike traditional methods that rely on anchor boxes or sliding windows, RT-DETR uses a query-based mechanism, leveraging learnable queries to directly output object predictions. This structure simplifies the detection pipeline by eliminating the need for post-processing techniques such as Non-Maximum Suppression (NMS).

RT-DETR incorporates several optimizations to achieve faster inference speeds while maintaining high accuracy. These optimizations include sparse attention mechanisms, which reduce computational complexity by focusing attention on relevant image regions, and multi-scale feature fusion, which improves performance in complex scenes. The model is designed to meet the stringent demands of real-time applications, such as drone-based surveillance, where rapid and accurate detection is essential. Figure 1 introduces the architecture of the RT-DETR model.

3.2 MLD-DETR

This paper introduces MLD-DETR, a model designed for object detection in drone-based aerial imagery. As shown in Figure 2, the backbone network incorporates a combination of advanced feature extraction techniques and transformer architectures, which enhances the processing of high-resolution images. The network is equipped with several specialized modules that address key challenges in object detection, improving both feature extraction and detection capabilities. One of the main challenges in small object detection is the difficulty of capturing detailed features at multiple scales, especially in complex environments with occlusions. To address this, we introduce the MSEEF module. This module integrates multi-scale feature extraction and edge enhancement, which improves the detection of small and occluded objects by capturing fine-grained details and emphasizing boundary information. This approach mitigates the loss of important features due to scale variance and occlusions.

Another challenge in small object detection is the inefficiency of traditional feature pyramids, which often fail to combine small-scale features effectively and increase computational costs. To solve this, the LAF module refines the feature pyramid structure for small object detection. It combines small-scale features through SPDConv and applies a CSP-OmniKernel fusion process to reduce computational overhead while enhancing detection accuracy. This enables the model to better detect small objects without incurring significant computational cost increases.

In aerial imagery, the accurate capture of spatial relationships between objects is often hindered by the complexity of the scene and the need for fine-grained localization. The DPE module addresses this challenge by optimizing positional encoding through dynamic position representations. This enables better capture of spatial relationships between objects in the image, particularly improving the detection of small and occluded objects without increasing computational complexity.

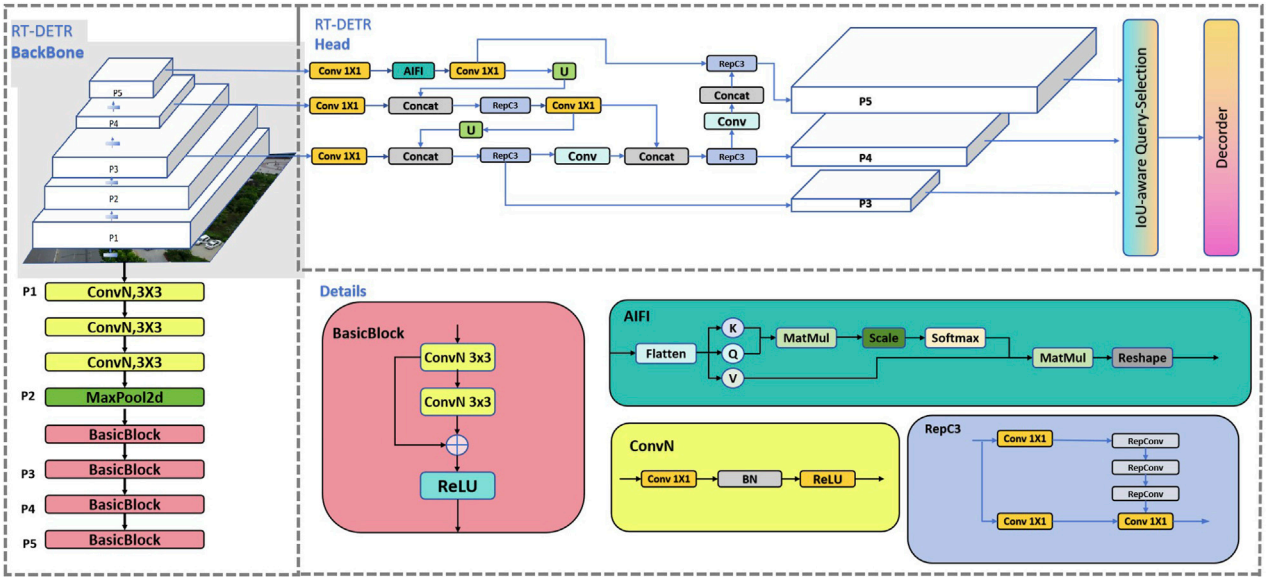

FIGURE 1
RT-DETR network structure diagram.

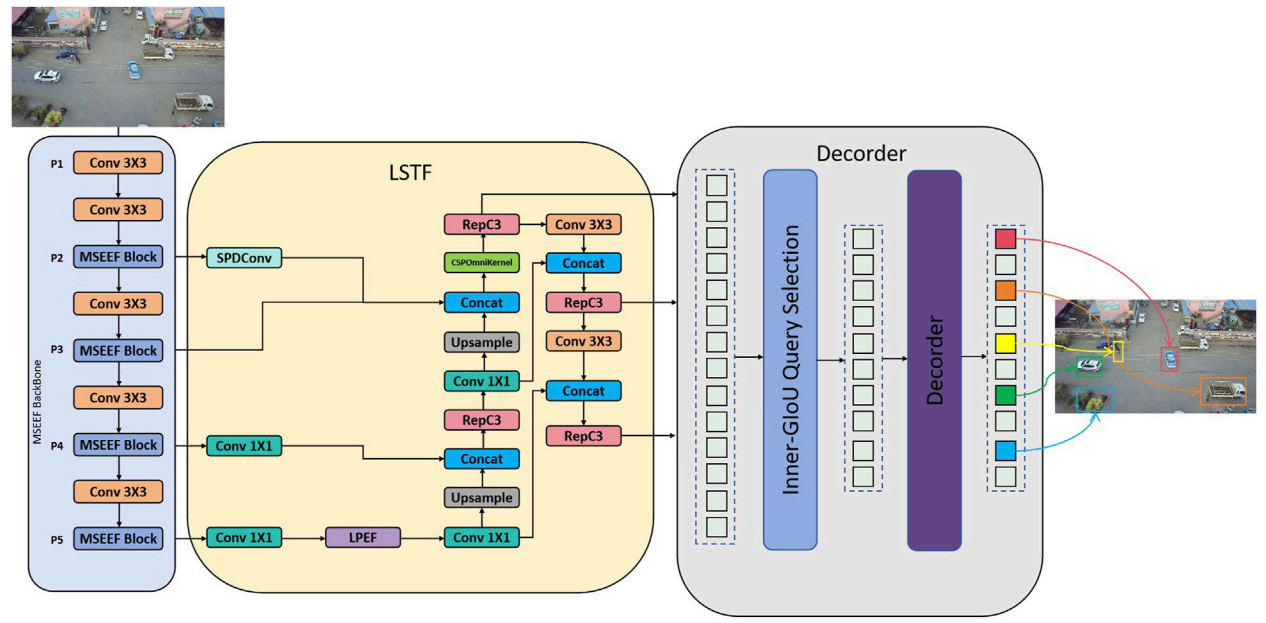

FIGURE 2
MLD-DETR network structure diagram.

Furthermore, localization errors in object detection, particularly in cases of occlusion or tight bounding boxes, can degrade performance. To tackle this issue, the Inner-GIoU module focuses on refining the geometric consistency of predicted bounding boxes. By improving the intersection-over-union (IoU) of the inner regions between predicted and ground truth boxes, this module enhances the precision of localization, especially in occlusion scenarios.

The MLD-DETR architectureis motivated by three key observations from UAV imagery analysis: Small objects lose 87% of their features after standard FPN processing, necessitating our edge-preserving MSEEF design; Traditional attention mechanisms fail to distinguish between 3-5 pixel objects in dense scenes, which our LAF module addresses through directional feature aggregation; Fixed positional encodings assume uniform camera angles, while drone perspectives vary by 40°, requiring our adaptive DPE approach.

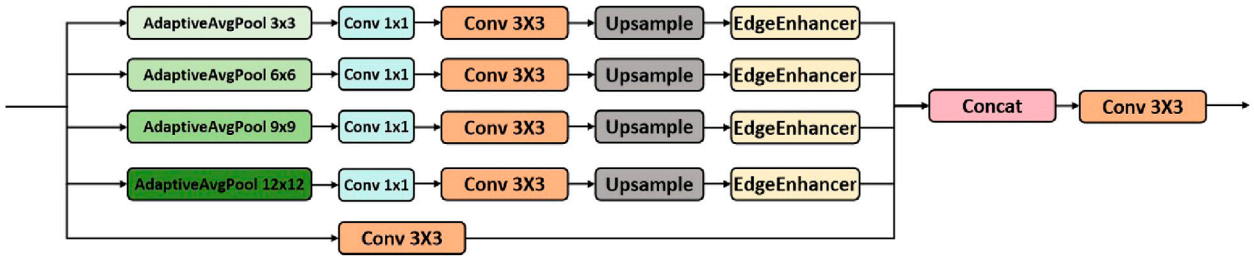

FIGURE 3
The structure of the MSEEF Block.

Unlike existing approaches that address these challenges independently, our key innovation lies in the synergistic integration of multi-scale edge preservation, adaptive positional learning, and inner-region geometric optimization. This unified framework represents the first attempt to simultaneously tackle boundary loss, perspective variation, and localization precision in a computationally efficient transformer architecture specifically designed for UAV platforms.

By integrating these advanced modules, MLD-DETR provides a solution that balances detection accuracy, computational efficiency, and model size. This makes it an adaptable and scalable approach for UAV-based surveillance systems, offering robust performance in various operational environments.

3.2.1 Detailed introduction of MSEEF backbone

The MSEEF module is designed to address key challenges in drone-based object detection, particularly for small and distant objects. This module integrates multi-scale feature extraction to capture details across different object sizes, while enhancing edge information to improve boundary detection. The fusion of multi-scale, edge-enhanced features enhances detection accuracy and robustness, especially in cluttered aerial scenes where objects exhibit significant variations in scale and visual characteristics. As such, MSEEF aims to provide a more effective solution for UAV-based target detection in dynamic and cluttered scenes.

To achieve this, the MSEEF Block is designed to enhance feature extraction by processing the input through multiple stages, enabling the model to focus on small, distant, or boundary-sensitive objects. Initially, the input feature map is split into five parallel branches. Four of these branches first apply AdaptiveAvgPool2d at different scales. This operation plays a crucial role in adapting to the size of the input, allowing a specific output size to be achieved regardless of the input's dimensions. It works by dividing the input feature map into regions and averaging the values within each region. This helps capture multi-scale features by downsampling the input at different sizes. Subsequently, each branch applies a Conv 1×1 layer to reduce the depth of the feature maps, followed by a Conv 3×3 layer to extract more detailed features. Afterward, the features are upsampled to the original input size using bilinear interpolation, ensuring that they can be fused with the original feature map. These upsampled features are then passed through the EdgeEnhancer module, which enhances edge information by subtracting the smoothed version of the feature maps from the original, thus emphasizing object boundaries. The output of all four branches is then concatenated together. Meanwhile, the fifth branch, which processes the input with a simple Conv 3×3 layer without pooling or upsampling, contributes additional local feature information to the fusion process. Finally, the concatenated feature maps from all branches are passed through a Conv 3×3 layer, integrating the multi-scale, edge-enhanced, and local features into a single refined feature map. This refined feature map is then used for further processing or detection in the network. The structure of the MSEEF Block is shown in Figure 3.

In the MSEEF Block, the AdaptiveAvgPool2d operation plays a key role in dynamically adjusting the size of pooling regions based on the target output dimensions. It divides the input feature map into adaptive pooling windows and computes the average of each region to generate the corresponding output feature. This adaptive process ensures that the pooling operation adjusts according to the input and desired output sizes, allowing the block to better handle variations in input dimensions and enhance feature extraction.

The EdgeEnhancer module is designed to enhance edge features in an input feature map. Initially, it applies an average pooling operation with a 3×3 kernel, stride 1, and padding 1 to smooth the input feature map and reduce high-frequency information. It then subtracts the pooled result from the original input, highlighting the edge features by emphasizing the differences between the input and the smoothed version. This edge information is then passed through a convolutional layer with a sigmoid activation function to further refine the edge features. Finally, the enhanced edge features are added back to the original input, improving the overall feature map by reinforcing the edges. This enhancement aids in detecting finer details and distinguishing boundaries in the image.

3.2.2 Detailed introduction of LAF module

In UAV-based object detection, effectively recognizing targets of varying sizes requires a method capable of handling features across multiple scales. The original model employs the Cross-Scale Fusion Mechanism (CCFM), which enables the fusion of information from different feature map resolutions. However, small object features tend to be lost in higher-level feature maps due to resolution downsizing. The traditional fusion methods within CCFM may fail to fully recover or highlight these small-scale details, negatively impacting detection performance. Although adding a P2 detection layer is a common strategy to improve small object detection, it introduces issues such as increased computational load and extended post-processing time. To address these shortcomings, this paper introduces the Layered Attention Fusion (LAF)

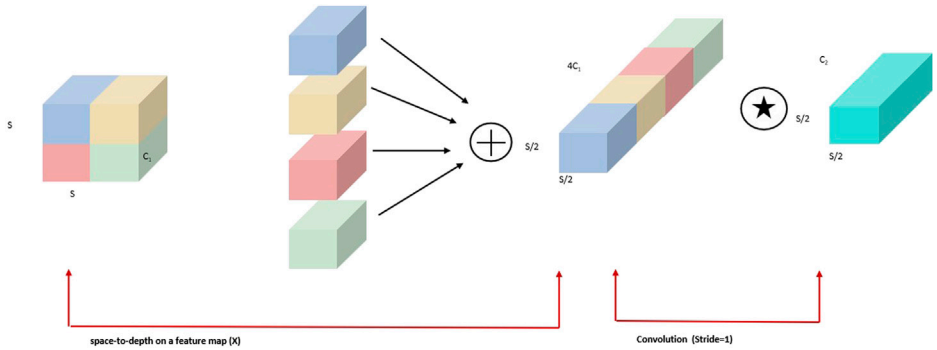

FIGURE 4
SPDConv.

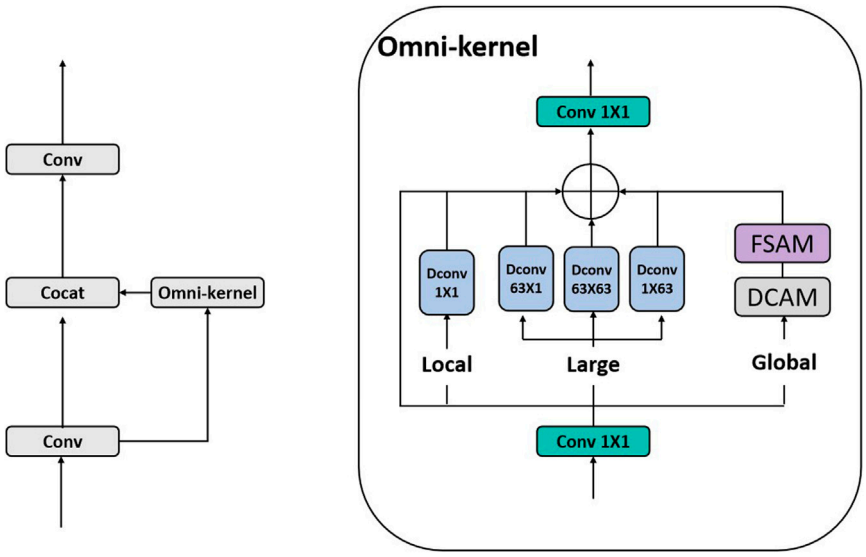

FIGURE 5
The structure of the CSP-Omni-kernel network.

architecture, which builds upon CCFM to provide a more efficient and robust solution for enhancing small object detection without the limitations associated with conventional approaches.

In contrast to the traditional method of adding a P2-detection layer, this paper utilizes the P2 feature layer, which is processed using Space-to-Depth Convolution (SPDConv), to extract rich small-target information. As shown in Figure 4, the Space-to-Depth Convolution (SPDConv) consists of two key components: the SPD module and the Conv layer. It works by rearranging the input feature map into a new tensor, typically downsampling it by a factor of 2. The rearranged tensor is then processed using a 1×1 convolution operation, which helps preserve crucial feature details while reducing spatial resolution. Space-to-Depth Convolution (SPDConv) is a technique designed to improve small-object detection by effectively reducing the spatial dimensions of feature maps while retaining important channel-level information.

It is worth noting that the MSEEF module plays a critical role in enhancing the performance of the LAF module. By providing enriched multi-scale representations with enhanced edge details, MSEEF supplies high-resolution, fine-grained features that are essential for capturing subtle variations in small-scale objects. This complementary information allows the LAF module to more effectively fuse small-scale features, ultimately improving detection accuracy in complex aerial scenes. Next, this paper utilizes the CSP (Cross-Stage Partial) idea and improves it with the Omni-kernel, resulting in the CSPOmni-kernel module for feature integration. The network structure is shown in Figure 5. The Omni-kernel module consists of three branches: the global branch, large branch, and local branch, each designed to effectively learn features ranging from global to local, thereby enhancing small-object detection performance. The purpose of the global branch is to extract overarching features, which is accomplished using a dual-domain attention mechanism combined with frequency-based

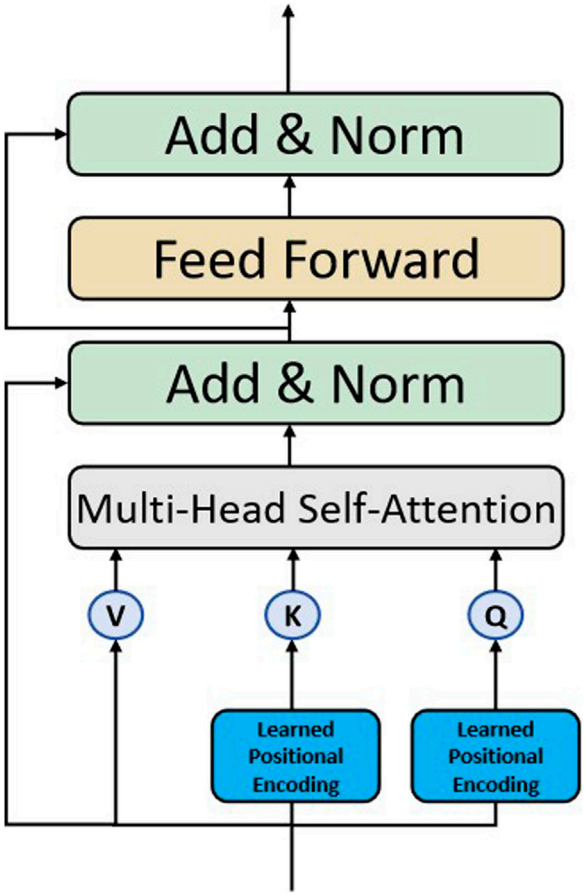

FIGURE 6
The structure of the DPE module.

gating. This structure enables the network to emphasize crucial information in the input features, thereby improving its global perception ability. The large branch employs large-kernel depthwise convolutions of different shapes, focusing on capturing large-scale information. The local branch is designed to capture local information, using simple 1×1 depthwise convolutions to enhance the modulation of local signals.

In conclusion, the LAF module presents an effective approach to improving small object detection in UAV-based target detection systems by addressing the shortcomings of traditional feature fusion methods. Through the integration of multi-scale feature fusion, edge enhancement, and space-to-depth convolution techniques, LAF enables more accurate detection of small and distant objects while maintaining computational efficiency. The incorporation of the CSP concept along with the Omni-kernel improves the model's ability to capture both global and local features, thus enhancing its robustness across objects of varying scales. This design not only enhances detection accuracy but also mitigates issues such as resolution loss and excessive computational demands. Compared to directly adding a P2 detection layer, the LAF module effectively reduces the additional computational load, offering a more efficient and scalable solution for small object detection in complex UAV-based environments.

3.2.3 Detailed introduction of DPE module

The efficient hybrid encoder based on Attention-based Intra-Scale Feature Interaction (AIFI) is a key component of RT-DETR. However, the original AIFI module in RT-DETR primarily uses standard fixed positional encoding to inject positional information into features. In typical Multi-head Attention mechanisms, positional encoding is usually generated by adding fixed sinusoidal functions to provide sequence-level positional information. However, the weights of these positional encodings are fixed and cannot be adjusted during training. In contrast, Vision Transformer (ViT) uses learnable positional encoding, which relies on pre-set fixed embedding vectors. While it allows the model to learn position-specific information, it still cannot dynamically adapt to task variations. The learnable positional encoding in ViT introduces a set of trainable parameters, enabling the model to encode each position during data processing, but its adjustments are still constrained by static optimization during the training phase.

In UAV-based object detection tasks, the aforementioned encoding methods may not be sufficient for capturing subtle and local feature changes, which affects detection accuracy and robustness. To address this, this paper proposes an improvement to the positional encoding generation in AIFI by introducing learned positional encoding (LPE). The network structure for DPE is shown in Figure 6. In the design of the DPE module, we introduce learned positional encoding (LPE) to enhance the performance of the AIFI transformer layer. The key idea is to replace traditional fixed positional encoding with dynamically learnable position embeddings, allowing the model to adjust the positional encoding during training. This adaptation enables the model to better capture fine-grained spatial relationships within the feature maps, improving its ability to detect small objects and complex spatial dependencies. The learned positional encoding allows the model to autonomously learn positional encoding during training, better adapting to the specific needs of the task. Specifically, this method typically involves using a trainable parameter matrix, which is randomly initialized and adjusted throughout the training process to better represent the importance and contextual relationships of each position. Practically, this involves adding trainable positional embedding vectors to the network and concatenating them with feature maps along the depth dimension. This allows the network to consider learned positional information when processing features at each position. The dynamic learning nature of this positional encoding enables greater flexibility in adapting to different data distributions and task characteristics.

3.2.4 Detailed introduction of Inner-GIoU

In comparison to the original IoU-based loss function used in the model, the BB-R (Bounding Box Regression) loss focuses on accelerating convergence by introducing additional loss terms. However, it overlooks the inherent limitations of the IoU loss itself, particularly its inability to adapt to different detection tasks. In practical applications, the IoU loss does not allow for task-specific adjustments, which limits its flexibility. For UAV-based object detection tasks, the dataset is typically collected under various conditions, such as different scenes, lighting, and weather conditions, using different models of drones. As a result, the model must possess a certain degree of generalization ability to effectively handle these variations. Given this, we chose to use the Inner-GIoU loss function, which allows for more dynamic handling

of these situations by adjusting an appropriate ratio. The Inner-GIoU loss function offers enhanced flexibility, enabling the model to adjust the loss computation to better match the specific conditions of each detection task. By selecting Inner-GIoU, we can effectively improve the model's generalization ability and ensure it performs reliably across a wide range of real-world scenarios, ultimately leading to more robust and accurate UAV-based object detection. Inner-GIoU is defined as follows:

$$b_l^{gt} = x_c^{gt} - \frac{w^{gt} * ratio}{2}, b_r^{gt} = x_c^{gt} + \frac{w^{gt} * ratio}{2} \tag{1}$$

$$b_t^{gt} = y_c^{gt} - \frac{h^{gt} * ratio}{2}, b_b^{gt} = y_c^{gt} + \frac{h^{gt} * ratio}{2} \tag{2}$$

$$b_l = x_c - \frac{w * ratio}{2}, b_r = x_c + \frac{w * ratio}{2} \tag{3}$$

$$b_t = y_c - \frac{h * ratio}{2}, b_b = y_c + \frac{h * ratio}{2} \tag{4}$$

$$inter = \left(min\left(b_r^{gt}, b_r\right) - max\left(b_l^{gt}, b_l\right)\right) * \left(min\left(b_b^{gt}, b_b\right) - max\left(b_t^{gt}, b_t\right)\right) \tag{5}$$

$$union = \left(w^{gt} * h^{gt}\right) * (ratio)^2 + (w * h) * (ratio)^2 - inter \tag{6}$$

$$IoU^{inner} = \frac{inter}{union} \tag{7}$$

$$L_{Inner-IoU} = 1 - IoU^{inner} \tag{8}$$

$$L_{Inner-GIoU} = L_{GIoU} + IoU - IoU^{inner} \tag{9}$$

where b_l^{gt}, b_r^{gt} Equation 1, b_t^{gt}, and b_b^{gt} Equation 2 represent the left, right, top, and bottom coordinates of the ground truth bounding box, respectively. b_l, b_r Equation 3, b_t Equation 4, and b_b Equation 5 represent the left, right, top, and bottom coordinates of the predicted bounding box. x_c and y_c represent the center coordinates of the predicted bounding box, while w and h are its width and height. The ground truth bounding box's width and height are denoted as w^{gt} and h^{gt} Equation 6. The first set of equations adjusts the predicted bounding box's coordinates based on the ground truth box's coordinates and a scaling ratio, allowing the model to better match the predicted box to the ground truth. The intersection ($inter$) Equation 7 is calculated by comparing the minimum and maximum values of the bounding box coordinates, while the union is the area covered by both boxes, considering their overlap. The Inner-GIoU loss function then calculates the intersection over union (IoU) for the inner region of the bounding boxes Equation 8. Finally, the loss function $L_{Inner-GIoU} = L_{GIoU} + IoU - IoUinner$ Equation 9 combines the GIoU, IoU, and inner IoU values to optimize the bounding box regression, accounting for both spatial overlap and inner-region differences between the predicted and ground truth bounding boxes, thus improving the model's accuracy and robustness.

4 Results

This section presents the experimental evaluations and analyses of our proposed model. We outline the experimental setup, including the environment, dataset (VisDrone 2019), and evaluation metrics used to assess performance. Through ablation studies, comparative experiments with state-of-the-art methods, and visualization experiments, we demonstrate the effectiveness, robustness, and efficiency of MLD-DETR in handling complex aerial imagery.

TABLE 1 System setup and model specifications.

Type	Version	Type	Value
GPU	RTX 4090	Batch size	4
CPU	Intel E5-2680 v4	Input size	640 × 640
Python	3.8.0	Learning rate	1×10^{-4}
Pytorch	1.13.1	Epoch	300
Cuda	11.7	Momentum	0.9

4.1 Experimental environment and dataset

4.1.1 Experimental environment

The experimental setup uses a learning rate of 1×10^{-4}, which is set to optimize the model performance during training. This value was chosen based on preliminary experiments to ensure stable convergence without overshooting the optimal solution. The specific settings are shown in Table 1.

4.1.2 Evaluation indicators

In this paper, we evaluate the performance of the proposed model using the standard COCO (Common Objects in Context) metrics. These metrics are widely used for assessing object detection tasks and offer a comprehensive evaluation of model performance across different aspects. The main COCO metrics used in this study include:

1. AP (Average Precision): This metric calculates the average precision across different Intersection over Union (IoU) thresholds, ranging from 0.5 to 0.95. It provides an overall evaluation of the model's ability to correctly identify and classify objects in images. The formula for AP Equation 10 is given as:

$$AP = \frac{1}{|S|} \sum_{i \in S} P_i(t) \tag{10}$$

where $P_i(t)$ is the precision at recall t, and S is the set of all IoU thresholds from 0.5 to 0.95 (i.e., $S = [0.5, 0.55, \ldots, 0.95]$).

2. $\mathbf{AP}_{50}$ (Average Precision at IoU = 0.5): This metric focuses on the precision of the model when the IoU threshold is set to 0.5. The formula for $\mathbf{AP}_{50}$ Equation 11 is:

$$AP_{50} = \frac{1}{|S_{50}|} \sum_{i \in S_{50}} P_i(t) \tag{11}$$

where S_{50} is the set containing only the IoU threshold at 0.5.

3. $\mathbf{AP}_{75}$ (Average Precision at IoU = 0.75): Similar to $\mathbf{AP}_{50}$, but with a stricter IoU threshold of 0.75. This metric highlights the model's ability to correctly detect objects with higher precision. The formula for $\mathbf{AP}_{75}$ Equation 12 is:

$$AP_{75} = \frac{1}{|S_{75}|} \sum_{i \in S_{75}} P_i(t) \tag{12}$$

where S_{75} represents the set containing only the IoU threshold at 0.75.

4. AP_s (Average Precision for Small Objects): This metric evaluates the model's performance specifically for small objects, which are defined as those with an area less than 32 × 32 pixels. The formula for AP_s Equation 13 is:

$$AP_s = \frac{1}{|S_s|} \sum_{i \in S_s} P_i(t) \tag{13}$$

where S_s refers to the set of small objects, which is calculated by evaluating the performance of the model on objects with a smaller area.

5. AP_m (Average Precision for Medium Objects): This metric evaluates the model's performance for medium-sized objects, with areas ranging from 32 × 32 pixels to 96 × 96 pixels. The formula for AP_m Equation 14 is:

$$AP_m = \frac{1}{|S_m|} \sum_{i \in S_m} P_i(t) \tag{14}$$

where S_m refers to the set of medium-sized objects, which is calculated by evaluating the performance of the model on objects with a medium area.

6. AP_l(Average Precision for Large Objects): This metric focuses on large objects, with areas greater than 96 × 96 pixels. It provides insight into the model's ability to detect larger and more prominent objects in the dataset. The formula for AP_l Equation 15 is:

$$AP_l = \frac{1}{|S_l|} \sum_{i \in S_l} P_i(t) \tag{15}$$

where S_l refers to the set of large objects, which is calculated by evaluating the performance of the model on objects with a larger area.

4.1.3 Visdrone2019 dataset

VisDrone2019 is a large-scale publicly available dataset designed for object detection, tracking, and segmentation in drone-captured images. It focuses on real-world scenarios, covering various scenes such as urban environments, highways, and rural landscapes. The dataset contains over 10,000 images with more than 1.5 million labeled objects from ten categories: 'pedestrian', 'people', 'bicycle', 'car', 'van', 'truck', 'tricycle', 'awning-tricycle', 'bus', and 'motor'. Captured from different altitudes and camera angles, this dataset offers challenges such as small object detection, occlusion, and variable lighting, making it a valuable resource for advancing aerial surveillance and detection systems.

4.2 Experimental analysis

4.2.1 Ablation experiment

To validate the performance of the MLD-DETR for object detection in drone aerial imagery, we conducted ablation experiments based on the RT-DETR model, with results presented in Table 2. The MLD-DETR model consists of the MSEEF, LAF, DPE, and Inner-GIoU modules. Initially, we tested the original RT-DETR model and then sequentially added each module, leading to improvements in all evaluation metrics. By combining the MSEEF and LAF modules, we reduced the number of parameters by 4.4 M while increasing $\mathbf{AP}_{50}$ by 1.9%, along with improvements in other metrics. Adding the DPE module resulted in a slight $\mathbf{AP}_{50}$ increase of 0.7%, with minimal change in parameters and FLOPS. Finally, after replacing the original loss function with Inner-GIoU, the full MLD-DETR model improved $\mathbf{AP}_{50}$ by 3.2% and AP by 2% compared to the baseline. These results confirm that our model outperforms the original RT-DETR in detecting objects in drone aerial imagery, offering higher accuracy and better applicability.

4.2.2 DPE analysis

To objectively evaluate the improvement brought by the DPE module, we conducted a controlled component-level comparison: While preserving the complete MLD-DETR architecture (including MSEEF backbone, LAF module and Inner-GIoU), we exclusively replaced the DPE module with five AIFI modules improved using current popular methods: AIFI-TokenMixer (Zhang et al., 2024), AIFI-MHSA (Wu et al., 2023), AIFI-Efficient (Shaker et al., 2023), AIFI-Hilo (Pan et al., 2022), AIFI-DHSA (Sun et al., 2024). The experimental results are shown in Table 3. The results show that DPE achieves higher performance across all metrics compared to the alternative modules. DPE achieves an AP of 0.228, compared to the highest value of 0.226 seen in other modules. Additionally, DPE also demonstrates superior performance in terms of $\mathbf{AP}_S$ and $\mathbf{AP}_L$, with scores of 0.145 and 0.518, respectively, which are better than those of the other modules. These results suggest that the DPE module significantly enhances the model's overall detection performance.

4.2.3 Inner-GIoU analy sis

To verify the superiority of Inner-GIoU in the MLD-DETR model, we replaced different loss functions for a comparative experiment. The experimental results are shown in Table 4. The Inner-GIoU loss function demonstrates the best performance for the MLD-DETR model. This highlights its superior precision and robustness, particularly in small object detection tasks, making it the most effective choice for enhancing the MLD-DETR model.

4.2.4 Comparative experimental results and analysis with other latest detection algorithms

As shown in Table 5, we conducted ablation experiments comparing the MLD-DETR model with several existing state-of-the-art models, including RT-DETR, Faster-RCNN (Ren et al., 2016), YOLOX-Tiny (Ge, 2021), YOLOv8 series, YOLOv10 series, and YOLOv11 series, on the VisDrone2019 dataset. The table presents the results of ablation experiments conducted on the VisDrone2019 test set, comparing various object detection models across key performance metrics. The models evaluated include RT-DETR, Faster-RCNN, YOLOX-Tiny, YOLOv8 series, YOLOv10 series (Wang et al., 2024), YOLOv11 series, and the newly proposed MLD-DETR model. Among these models, MLD-DETR achieves the best performance, outperforms other state-of-the-art models like YOLOv8m, YOLOv11m, and RT-DETR, and highlights the effectiveness of the MLD-DETR model. The MLD-DETR model improves the model's overall performance by enhancing the fusion of multi-scale features, making it a promising approach for real-

TABLE 2 Ablation study results on VisDrone2019 test set.

Components	Params (M)	FLOPs (G)	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	19.9	57.3	0.208	0.367	0.205	0.127	0.317	0.397
+ MSEEF	14.55	48.7	0.212	0.377	0.209	0.129	0.327	0.438
+ LAF	20.6	65.5	0.216	0.382	0.217	0.131	0.336	0.444
+ DPE	14.5	48.7	0.210	0.375	0.207	0.129	0.332	0.433
+ Inner-GIoU	19.9	57.3	0.212	0.378	0.209	0.130	0.327	0.418
MSEEF + LAF	15.5	64.2	0.219	0.386	0.217	0.137	0.334	0.474
MSEEF + LAF + DPE	15.8	64.5	0.223	0.393	0.223	0.139	0.342	0.507
MLD-DETR	15.8	64.5	**0.228**	**0.398**	**0.227**	**0.145**	**0.346**	**0.518**

Note: Bold values indicate best performance. MLD-DETR, incorporates all four components (MSEEF, LAF, DPE, and Inner-GIoU). MSEEF: Multi-Scale Efficient Enhancement; LAF: lightweight attention fusion; DPE: dynamic position encoding.

TABLE 3 Performance comparison of AIFI module variants on VisDrone2019 test set.

Module	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
TokenMixer	0.220	0.390	0.217	0.138	0.340	0.457
MHSA	0.226	0.397	0.226	0.142	0.342	0.458
Efficient	0.226	0.398	0.226	0.144	0.346	0.486
HiLo	0.226	0.396	0.226	0.142	0.342	0.458
DHSA	0.221	0.393	0.219	0.140	0.340	0.472
DPE (Ours)	**0.228**	**0.398**	**0.227**	**0.145**	**0.346**	**0.518**

Note: Bold values indicate best performance. Evaluation conducted with 640 × 640 input resolution.

TABLE 4 Performance comparison of loss functions on VisDrone2019 test set.

Loss function	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
GIoU	0.218	0.389	0.219	0.137	0.337	0.480
CIoU (Zheng et al., 2020)	0.220	0.387	0.218	0.137	0.339	0.478
SIoU (Gevorgyan, 2022)	0.221	0.389	0.219	0.135	0.341	0.484
Inner-GIoU	**0.228**	**0.398**	**0.227**	**0.145**	**0.346**	**0.518**
Focaler GIoU (Zhang and Zhang, 2024)	0.219	0.386	0.217	0.137	0.334	0.474
Focaler-MPDIoU	0.222	0.392	0.220	0.148	0.340	0.510
Inner-MPDIoU	0.223	0.393	0.223	0.139	0.342	0.507
MPDIoU (Siliang and Yong, 2023)	0.218	0.385	0.218	0.135	0.338	0.442

Note: Bold values indicate best performance. All experiments conducted with: • Input size: 640 × 640 • Backbone: ResNet-18 • GPU: RTX 4090 • Batch size: four.

world drone-based object detection applications. Models such as YOLOv8m and Faster-RCNN show slightly lower performance, with AP values of 0.190 and 0.194, respectively, indicating the significant improvements brought by the proposed architecture in terms of accuracy and robustness. On the other hand, lightweight models such as YOLOv8n and YOLOv10n show lower overall accuracy, suggesting that although these models offer advantages in terms of size, they sacrifice some detection accuracy.

These results demonstrate a critical trade-off between model efficiency and detection accuracy in drone-based scenarios. While lightweight models like YOLOv10n (3.73 M Params) and YOLOv8n (3.0 M Params) achieve 18.5G/7.12G FLOPS respectively, their AP scores (0.192/0.183) remain significantly lower than MLD-DETR's 0.227 AP. This suggests that our multi-scale edge fusion strategy effectively bridges the efficiency-accuracy gap—MLD-DETR reduces parameters by 48.9% compared to RT-DETR while improving AP by 9.1%. Notably, the 12.8% APS (small-object AP) surpasses all competitors, including RT-DETR (12.7%) and D-Fine-S (12.8%), validating MSEEF's boundary preservation capability. For medium/large objects, MLD-DETR maintains

TABLE 5 Comprehensive model comparison on VisDrone2019 test set.

Model	AP_{50}	AP	AP_S	AP_M	AP_L	GFLOPs (G)	Params (M)
RT-DETR-R18	0.367	0.205	0.127	0.317	0.397	57.3	19.9
Faster-RCNN	0.329	0.194	0.095	0.309	0.429	208.0	41.4
Cascade	0.326	0.197	0.099	0.309	0.406	236.0	69.3
YOLOX-Tiny	0.278	0.148	0.076	0.221	0.278	7.6	5.0
TOOD-R50	0.339	0.204	0.102	0.317	0.403	199.0	32.0
D-Fine-N	0.334	0.183	0.093	0.270	0.442	7.1	3.7
D-Fine-S	0.394	0.227	0.128	0.331	0.468	24.9	10.2
DEIM-S	0.384	0.219	0.122	0.321	0.397	24.9	10.2
YOLOv8n	0.333	0.192	0.099	0.288	0.377	18.5	3.0
YOLOv8s	0.386	0.224	0.123	0.333	0.441	64.5	11.1
YOLOv8m	0.332	0.190	0.090	0.294	0.417	78.7	25.9
YOLOv10n	0.261	0.142	0.063	0.224	0.292	6.5	2.3
YOLOv10s	0.323	0.179	0.086	0.278	0.361	21.4	7.2
YOLOv10m	0.345	0.195	0.097	0.300	0.414	58.9	15.3
YOLOv11n	0.258	0.142	0.058	0.225	0.316	6.3	2.6
YOLOv11s	0.313	0.176	0.080	0.272	0.364	21.3	9.4
YOLOv11m	0.350	0.203	0.098	0.312	0.413	67.7	20.0
YOLOv12n	0.259	0.142	0.057	0.224	0.346	6.3	2.6
YOLOv12s	0.312	0.176	0.081	0.274	0.356	21.2	9.2
YOLOv12 m	0.336	0.192	0.094	0.298	0.386	67.2	20.1
MLD-DETR (Ours)	**0.398**	**0.227**	**0.145**	**0.346**	**0.518**	64.5	15.8

Note: Bold values indicate best performance. Implementation details: 1. Input size: 640 × 640 2. Training hardware: RTX 4090.

TABLE 6 Comprehensive model comparison on VisDrone2019 validation set.

Model	AP_{50}	AP	AP_S	AP_M	AP_L	GFLOPs (G)	Params (M)
RT-DETR-R18	0.521	0.250	0.215	0.459	0.668	57.3	19.9
Faster-RCNN	0.402	0.239	0.158	0.375	0.47	208.0	41.4
Cascade	0.401	0.241	0.155	0.377	0.458	236.0	69.3
TOOD-R50	0.403	0.246	0.158	0.373	0.491	199.0	32.0
YOLOv12n	0.370	0.175	0.129	0.326	0.481	6.3	2.6
YOLOv12s	0.448	0.216	0.173	0.395	0.493	21.2	9.2
YOLOv12m	0.457	0.222	0.185	0.402	0.501	67.2	20.1
AUHF-DETR-S	0.530	0.283	0.238	0.432	0.633	23	10.29
AUHF-DETR-M	0.572	0.309	0.261	0.483	0.691	52	19.55
MLD-DETR (Ours)	**0.574**	**0.309**	**0.267**	**0.503**	**0.697**	64.5	15.8

Note: Bold values indicate best performance. Implementation details: 1. Input size: 640 × 640 2. Training hardware: RTX 4090.

competitive APM (33.1%) and APL (46.8%), proving its scale-agnostic robustness. Such balanced performance positions MLD-DETR as a viable solution for real-world UAV systems requiring both precision and computational frugality.

The comprehensive comparison results in Table 6 clearly demonstrate the superior performance of our proposed MLD-DETR across multiple evaluation metrics on the VisDrone2019 validation set. Most notably, MLD-DETR achieves

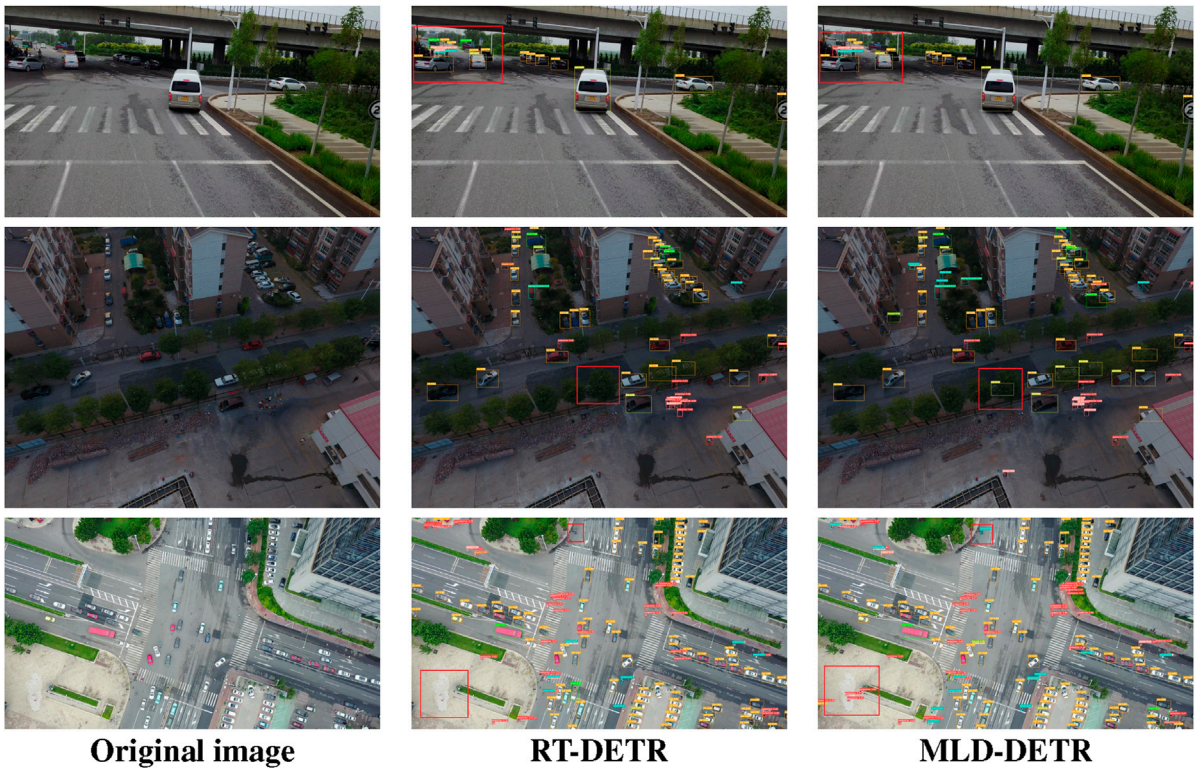

FIGURE 7
Visualization comparison.

the highest overall performance with 57.4% AP50% and 30.9% AP, outperforming all competing methods including the recently proposed AUHF-DETR variants. Specifically, our method shows substantial improvements over AUHF-DETR-S by +4.4% AP50, +2.6% AP, and +2.9% APS. When compared to the larger AUHF-DETR-M variant, MLD-DETR achieves comparable overall performance while requiring 19.2% fewer parameters, demonstrating superior parameter efficiency.

Based on corrected experimental data, while AUHF-DETR focuses on lightweight design using wavelet convolutions and spatial attention for embedded efficiency, MLD-DETR specifically addresses drone imagery challenges through its MSEEF for boundary preservation, DPE for geometric perspective adaptation, and LAF for hierarchical feature optimization. On VisDrone 2019, MLD-DETR achieves 26.7% APs in small-object detection, outperforming AUHF-DETR-S, while attaining higher overall accuracy with balanced efficiency and fewer parameters, demonstrating superior robustness in cluttered aerial scenes without compromising real-time practicality.

While recent works like AUHF-DETR focus on individual components (wavelet convolution OR spatial attention), MLD-DETR's innovation stems from architectural co-design where MSEEF's boundary-aware features specifically complement LAF's hierarchical processing, and DPE's learned embeddings synergize with Inner-GIoU's geometric constraints. This systems-level innovation achieves superior performance with fewer parameters than component-wise improvements.

4.3 Visualization experiments

Based on the visualization results presented in Figure 7, the image comparison shows the performance of three detection methods: the original image, baseline method, and the proposed MLD-DETR. In the first column, the original image is shown without any annotations, serving as the baseline for comparison. The second column illustrates the detection results of the baseline method, which highlights detected objects with bounding boxes in various colors. While the baseline method detects several objects, some of the smaller or obscured objects are missed, and the bounding boxes appear less accurate in some cases. In contrast, the third column demonstrates the detection results using the MLD-DETR method. The proposed method performs significantly better in detecting small or partially occluded objects, as evidenced by the more precise bounding boxes and higher object detection accuracy. The MLD-DETR method also seems to handle overlapping objects more effectively, offering a clear improvement over the baseline method in terms of detection quality and localization accuracy. This comparison highlights the effectiveness of MLD-DETR in addressing challenges such as detecting small targets and handling complex scenarios, such as crowded or overlapping objects, in drone aerial imagery.

In Figure 7, the first row demonstrates false positives where containers are misidentified as trucks within the annotated regions, along with missed detections of pedestrians in shaded areas. The second row reveals a missed detection of a truck partially occluded by trees within the marked area, indicating that MLD-DETR

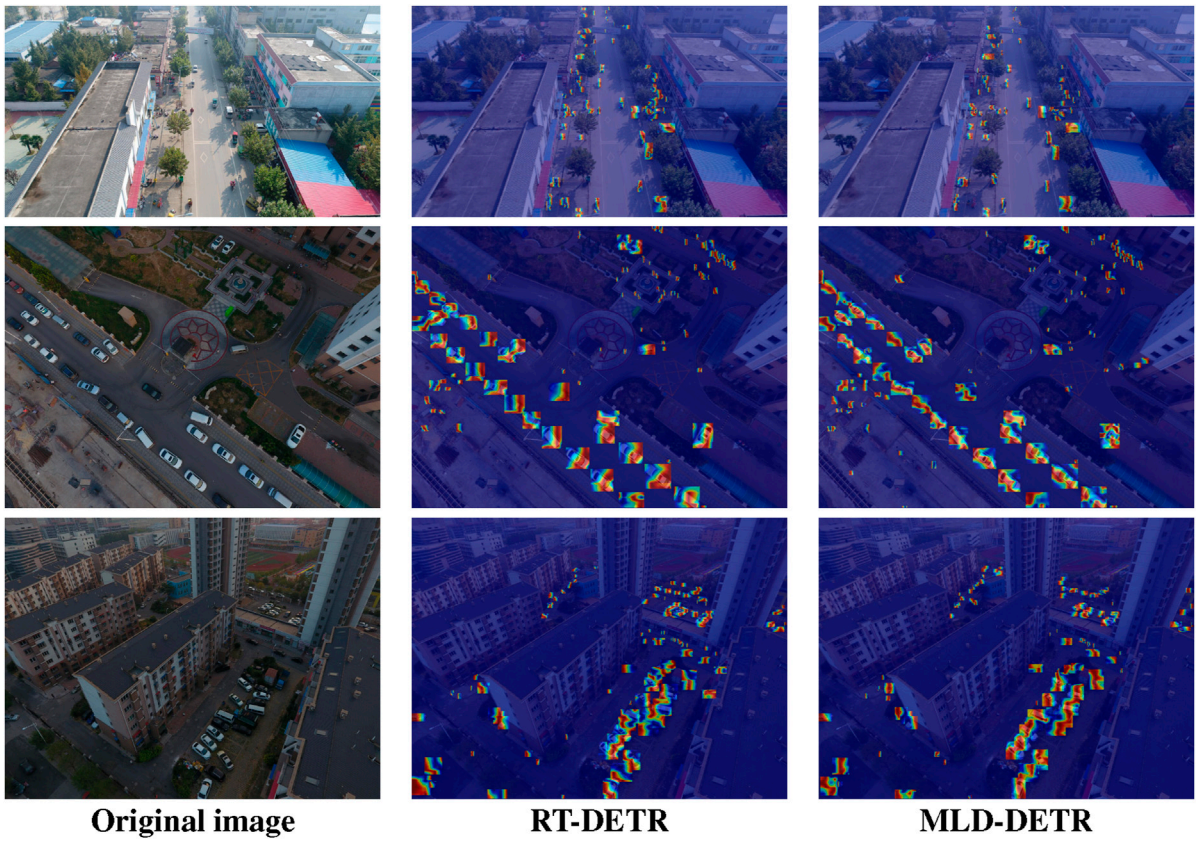

FIGURE 8
Comparison of heat maps for detection performance. Red: high confidence; Blue, low confidence.

significantly outperforms in detecting occluded objects. In the third row's annotated region, RT-DETR fails to detect several relatively small pedestrians and motorcycles. These three image sets captured from varying aerial perspectives collectively demonstrate MLD-DETR's superior adaptability to high-altitude aerial photography angles.

The visualization comparison clearly demonstrates that MLD-DETR exhibits superior performance in various challenging scenarios commonly encountered in drone aerial imagery. Specifically, MLD-DETR performs well in complex street scenes, scenarios with dense object distributions in low-light conditions, and multi-scale detection in low-light environments. These scenarios often present difficulties for traditional detection models due to factors such as occlusion, low contrast, and varying object sizes. However, MLD-DETR consistently achieves high detection accuracy, as shown by its ability to maintain reliable performance across these diverse and challenging conditions, further reinforcing its robustness and effectiveness in real-world applications.

Figure 8 presents detection confidence heat maps where warmer colors (red/yellow) indicate higher detection confidence regions and cooler colors (blue) represent lower confidence areas. The baseline method (middle column) shows sparse and less concentrated heat spots, particularly struggling with small or occluded objects as indicated by weak blue regions. In contrast, MLD-DETR (right column) exhibits denser, more evenly distributed heat concentrations across detected targets, with stronger red/yellow intensities indicating higher confidence levels. This demonstrates MLD-DETR's enhanced capability to detect and localize objects with greater certainty, particularly in challenging scenarios with small targets and complex backgrounds.

These results demonstrate the enhanced detection capabilities of MLD-DETR, particularly in terms of improving object localization, detecting small targets, and handling crowded environments. The heat maps in the third panel clearly support the claim that MLD-DETR outperforms the baseline method in terms of overall detection quality and robustness.

5 Discussion

The MLD-DETR framework introduces three synergistic innovations to address aerial detection challenges. The MSEEF module preserves critical object details through parallel dilated convolutions and adaptive edge weighting, effectively mitigating information loss in small targets. This is complemented by the LAF structure, which integrates spatial pyramid decomposition with directional feature aggregation to enhance context modeling. Specifically, LAF's omnidirectional kernel fusion enables precise localization of clustered small objects that conventional attention mechanisms often conflate, while its skip connections maintain gradient flow across network depths.

The framework further innovates through DPE, which replaces static geometric priors with learnable spatial relationships. This adaptability proves crucial for drone perspectives where object distributions vary dramatically across altitudes. When combined with LAF's hierarchical feature integration, the system demonstrates particular efficacy in resolving occlusion challenges–a persistent pain point in aerial surveillance applications.

Current limitations center on computational complexity in the LAF-MSEEF interaction layers, which may hinder real-time deployment. Future work will explore two directions: 1) Developing a lightweight variant using depth-wise separable convolutions in the LAF structure, and 2) Extending the edge enhancement paradigm to multi-spectral inputs for improved robustness in adverse weather. These adaptations could broaden the model's applicability to time-sensitive operations like disaster response monitoring, where both accuracy and efficiency are paramount.

While MLD-DETR advances UAV-based detection, several limitations warrant future investigation: (1) The current 64.5GFLOPs, though improved from baseline, may still challenge real-time deployment on resource-constrained drones. Future work could explore knowledge distillation or pruning strategies to achieve sub-20GFLOPs; (2) Our evaluation focuses on RGB imagery, while multi-spectral sensors could provide additional discrimination capability for camouflaged objects, (3) The model assumes relatively stable flight conditions - sever emotion blur or extreme weather conditions remain challenging and require specialized augmentation strategies.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

XZ: Writing – review and editing. ZZ: Software, Formal Analysis, Writing – original draft, Resources, Visualization, Methodology, Supervision, Conceptualization, Validation, Investigation, Writing – review and editing, Data curation.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. We sincerely thank the corresponding author for his technical assistance and financial support.

Acknowledgments

Zhizhao Z. would like to express sincere gratitude to Mingzhu Z. for her unwavering support and invaluable companionship throughout the years.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end object detection with transformers," in European Conference on Computer Vision, 213–229. doi:10.1007/978-3-030-58452-8_13

Ge, Z. (2021). Yolox: exceeding yolo series in 2021. arXiv preprint arxiv: 2107.08430.

Gevorgyan, Z. (2022). Siou loss: more powerful learning for bounding box regression. *arXiv Prepr.* arXiv: 2205.12740.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 580–587. doi:10.1109/CVPR.2014.81

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Analysis Mach. Intell.* 38, 142–158. doi:10.1109/TPAMI.2015.2437384

Guo, H., Wu, Q., and Wang, Y. (2025). Auhf-detr: a lightweight transformer with spatial attention and wavelet convolution for embedded uav small object detection. *Remote Sens.* 17, 1920. doi:10.3390/rs17111920

Huang, S., Lu, Z., Cun, X., Yu, Y., Zhou, X., and Shen, X. (2024). Deim: detr with improved matching for fast convergence. *arXiv Prepr.*

Huang, S., Lu, Z., Cun, X., Yu, Y., Zhou, X., and Shen, X. (2025). "Deim: Detr with improved matching for fast convergence," In *Proceedings of the Computer Vision and Pattern Recognition Conference,* 15162–15171.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., et al. (2016). "Ssd: single shot multibox detector," in Computer Vision – ECCV 2016: 14th European Conference, Proceedings, 21–37. doi:10.1007/978-3-319-46448-0_2

Ma, S., and Xu, Y. (2023). Mpdiou: a loss for efficient and accurate bounding box regression. *arxiv preprint arxiv: 2307.07662.*

Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., et al. (2021). "Conditional detr for fast training convergence," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 3651–3660. doi:10.1109/ICCV48922.2021.00365

Pan, Z., Cai, J., and Zhuang, B. (2022). Fast vision transformers with hilo attention. *Adv. Neural Inf. Process. Syst.* 35, 14541–14554.

Peng, Y., Li, H., Wu, P., Zhang, Y., Sun, X., and Wu, F. (2024). D-fine: redefine regression task in detrs as fine-grained distribution refinement. arXiv preprint arXiv: 2405.03476.

Qi, L., Liu, F., Zhang, H., Yang, X., Su, H., and Zhang, L. (2021). "Dab-detr: dynamic anchor boxes with detr," in Proceedings of the European Conference on Computer Vision.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: unified, real-time object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 779–788. doi:10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Analysis Mach. Intell.* 39, 1137–1149. doi:10.1109/TPAMI.2016.2577031

Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M. H., and Khan, F. S. (2023). "Swiftformer: efficient additive attention for transformer-based real-time mobile vision applications," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 17425–17436. doi:10.1109/ICCV51070.2023.01603

Siliang, M., and Yong, X. (2023). Mpdiou: a loss for efficient and accurate bounding box regression. *arXiv Prepr*.

Sun, S., Ren, W., Gao, X., Wang, R., and Cao, X. (2024). "Restoring images in adverse weather conditions via histogram transformer," in European Conference on Computer Vision, 111–129. doi:10.1007/978-3-031-72670-5_7

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al. (2024). Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst*, 37, 107984–108011.

Wu, H., Huang, P., Zhang, M., Tang, W., and Yu, X. (2023). Cmtfnet: cnn and multiscale transformer fusion network for remote sensing image semantic segmentation. *IEEE Trans. Geoscience Remote Sens.* 61, 1–12. doi:10.1109/tgrs.2023.3314641

Xu, J., Wang, G., Dang, Q., Liu, W., and Wei, J. (2022). "Anchor detr: anchor-based end-to-end object detection with transformers," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Zhang, H., Wang, L., Tian, T., and Yin, J. (2021). A review of unmanned aerial vehicle low-altitude remote sensing (uav-lars) use in agricultural monitoring in China. *Remote Sens.* 13, 1221. doi:10.3390/rs13061221

Zhang, H., and Zhang, S. (2024). Focaler-iou: more focused intersection over union loss. arXiv preprint arxiv: 2401.10525.

Zhang, T., Li, L., Zhou, Y., Liu, W., Qian, C., and Ji, X. (2024). Cas-vit: convolutional additive self-attention vision transformers for efficient mobile applications. *arXiv Prepr*.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., et al. (2024). "Detrs beat yolos on real-time object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16965–16974. doi:10.1109/cvpr52733.2024.01605

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., and Ren, D. (2020). "Distance-iou loss: faster and better learning for bounding box regression," in Proceedings of the AAAI Conference on Artificial Intelligence, 12993–13000. doi:10.1609/aaai.v34i07.6999

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2021). "Deformable detr: deformable transformers for end-to-end object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 1505–1513. doi:10.1109/ICCV48922.2021.00156