# LME-DETR: Lightweight and Multi-Scale Feature-Enhanced End-to-End Object Detection for Aerial Images

Feifei Xu[1*], Yu Xie[1*], Dongyang Li[1†], Luobin Huang[1], Zhihao Guo[2]
[1]School of Computer Science and Technology, Shanghai University of Electric Power, China
[2]School of Computer Science and Technology, University of Technology Sydney, Australia
xufeifei@shiep.edu.cn xieyusuep@mail.shiep.edu.cn dongyangli.ldy@shiep.edu.cn
luobinhuang043@gmail.com zhihao.guo@uts.edu.au

## Abstract

UAV-based object detection faces a fundamental challenge: achieving high accuracy on dense, tiny, and multi-scale objects while adhering to the stringent computational constraints of airborne platforms. To address this, we propose LME-DETR, a lightweight end-to-end Transformer detector co-designed for aerial perception and efficient deployment. LME-DETR introduces three key innovations. The Lightweight Receptive Enhancement Network expands effective receptive fields with near-zero inference overhead through reparameterizable dilated convolutions and channel-spatial attention, significantly enhancing contextual awareness for small objects. The Scale-Robust Feature Interaction Module (SRFIM) explicitly models cross-scale dependencies via dynamic normalization and multi-kernel depthwise convolutions, enabling adaptive fusion of multi-granularity features and improving robustness to extreme scale variations. The Exponential Moving Average Slide Varifocal Loss (EMA-SVFL) adaptively reweights hard, low-quality small-object samples by tracking their difficulty evolution during training, effectively countering the dominance of easy negatives in cluttered scenes. Extensive experiments on the VisDrone and UAVVaste benchmarks show that LME-DETR outperforms baselines in both accuracy and efficiency, while reducing parameters by 30% and GFLOPs by 19%, demonstrating a compelling balance of accuracy, efficiency, and real-time deployability for UAV applications.

Keywords: UAV object detection, Transformer, lightweight detector, Multi-scale feature interaction, Receptive field enhancement, Adaptive loss function

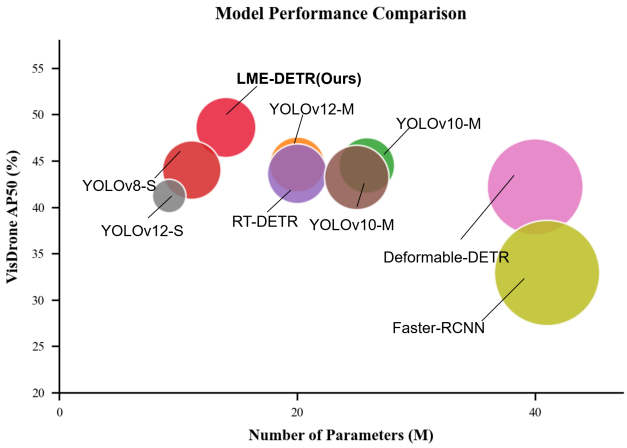


Figure 1. Model performance comparison on VisDrone dataset. The bubble size represents computational complexity (GFLOPs). Our method LME-DETR achieves the best accuracy with relatively low parameters and computational cost.

## 1. Introduction

Along with the rapid development of unmanned aerial vehicle (UAV) technology and the significant improvement in remote sensing image quality, object detection from the UAV perspective has become increasingly important in various application scenarios, such as environmental monitoring, traffic management, and disaster relief. However, due to the complexity of UAV-captured scenes, object detection in this domain faces numerous challenge [1]. These include significant variations in object scales and, most critically, the dense distribution of small targets, which are often defined as objects occupying an area of less than 32×32 pixels. Such targets, prevalent in aerial imagery, push the limits of current detection methods.

Typically, small objects occupy only a few pixels, and their semantic representation heavily relies on the

*These authors contributed equally to this work.
†Corresponding author.

surrounding contextual information [2]. If the model fails to capture sufficient semantic context within its receptive field, it is prone to missing or misclassifying the targets. Therefore, expanding the receptive field and enhancing the ability to model long-range dependencies are critical for small object detection [3]. In particular, constructing a wider contextual perception at shallow stages helps the model acquire complete semantic information early on, thereby enhancing the discriminability of small objects.

Existing strategies for enlarging the receptive field can be broadly categorized into three representative types. Dilated Convolution[4, 5, 6], for instance, expands the receptive field by introducing a dilation rate, offering advantages such as low parameter overhead and implementation simplicity. However, this method is prone to the gridding effect, resulting in discontinuous feature sampling and consequently weakening the representation of small objects. Multi-branch structure integrates receptive fields of various sizes by paralleling convolutional kernels with different scales[7, 8, 9]. Although it is effective in addressing scale variation, its structure complexity and high computational cost hinder its applicability in lightweight or real-time detection scenarios. Feature Pyramid Networks (FPN)[10, 11, 12, 13], and their variants introduce multi-scale semantics through cross-layer feature fusion and have become fundamental components in modern detection frameworks. However, FPN essentially functions as a feature aggregation mechanism and does not fundamentally expand the receptive field of a single layer. In particular, its capacity to enhance semantic representation at shallow stages remains limited.

In terms of long-range dependency modeling, attention mechanism is mainly employed. DEtection TRansformer(DETR) is designed for the first fully end-to-end, Transformer-based object detector, eliminating manual operation like Non-Maximum Suppression (NMS). Its highlight lies in the encoder's self-attention mechanism, achieving global context modeling by directly associating all feature positions within an image. However, the original DETR and its early variants have difficulty in slow convergence and high computational complexity. Especially, the quadratic complexity of global self-attention is particularly challenging for the high-resolution feature maps essential for small object detection, making these models ill-suited for real-time applications on resource-constrained UAV platforms.

Therefore, it is necessary to design a lightweight detection framework that can effectively enlarge the receptive field and enhance the ability to model long-range dependencies, thereby helping shallow-layer semantic modeling, while maintaining computational efficiency. To this end, we propose an efficient detection Transformer framework[14] tailored for UAV images, named LME-DETR. Compared with the baseline RT-DETR[15], LME-DETR features fewer training parameters and reduced computational overhead. Extensive experiments conducted on typical UAV image benchmark datasets, such as VisDrone-2019[16] and UAVVaste[17], demonstrate that LME-DETR significantly outperforms existing state-of-the-art DETR series models in terms of the trade-off between accuracy and computational efficiency across various model scales. Our main contributions are summarized as follows:

- We propose the Lightweight Receptive Enhancement Network, a novel architecture that achieves superior object perception through a synergism of reparameterizable dilated convolutions for broad contextual modeling and the attention-based mechanism for precise feature focusing.
- We present the Scale-Robust Feature Interaction Module to counteract feature degradation within Transformer encoders, achieving better feature preservation through a synergistic combination of a dynamic normalization strategy and multi-scale semantic fusion.
- We define a novel loss function, Exponential Moving Average Slide Varifocal Loss, which employs a sliding average regulation mechanism to adaptively shift the training focus and strengthen the optimization for hard-to-match small object samples.
- Finally, comprehensive comparisons with baselines and representative models designed for real-time aerial image detection demonstrate that our proposed LME-DETR achieves a superior balance between detection accuracy and computational efficiency, consistently delivering high-precision results while meeting real-time processing requirements.

## 2. Related Work

Considering the unique deployment environment of UAV platforms, the UAV-OD Unmanned Aerial Vehicle Object Detection(UAV-OD) task encounters challenges such as dense small objects, severe occlusion, and complex backgrounds, which makes demand of balancing model inference speed and computational resource constraints. Therefore, designing a detection model that maintains high accuracy while achieving

real-time performance and lightweight characteristics has attracted many researchers' attention.The mainstream methods can be broadly classified into three categories.

Two-stage methods, exemplified by Faster R-CNN [18], follow a region-based paradigm known for high accuracy. This framework typically involves a Region Proposal Network (RPN) generating candidate regions, followed by feature extraction for subsequent classification and regression.

The advantages of two-stage detectors lie in their precise localization and robustness. By decoupling region proposal from classification and utilizing RoI pooling, they focus computation on promising regions to extract rich features, ensuring high accuracy even for small or occluded objects [19].

However, these benefits come at the expense of efficiency. The separate proposal generation and complex inference pipeline result in high latency and resource consumption, making them less suitable for deployment on resource-constrained platforms like embedded UAV systems.

However, these benefits come at the expense of computation efficiency. The separate proposal generation and feature refinement result in high latency and resource consumption, making two-stage frameworks less suitable for deployment on platforms with limited computational power, especially embedded UAV systems. Moreover, their inference pipelines are relatively complex and difficult to optimize for real-time performance.

One-stage methods typically perform intensive predictions on feature maps at multiple scales, enabling effective detection of multi-scale objects. The YOLO series, for example, divides the input image into fixed grids for prediction and is widely adopted on edge devices due to its compact design and fast inference speed. Moreover, RetinaNet[20] addresses the issue of class imbalance by introducing Focal Loss, which enhances the learning capability for hard samples such as small objects.

One-stage detectors generally rely on predefined anchor boxes and post-processing techniques such as Non-Maximum Suppression (NMS), which limit their flexibility and adaptability in dynamic aerial environments. These limitations become particularly evident when dealing with densely distributed small objects or occlusions, where the network often struggles to capture discriminative features, resulting in frequent false positives and missed detections. Furthermore, the anchor-based matching mechanism tends to generalize poorly across varying object scales, imposing a bottleneck on model performance.

DETR-based methods leverage the Transformer architecture to model long-range dependencies and global contextual relationships, enabling end-to-end object detection without predefined anchors or post-processing like Non-Maximum Suppression. Detection Transformer(DETR) [21] was the first to formulate object detection as a direct set prediction problem, significantly simplifying the pipeline and distinct from traditional anchor-based approaches.

Despite its concise concept, DETR suffers from slow convergence and poor performance on small objects, largely due to its reliance on bipartite matching and insufficient exploitation of multi-scale features. To address this, Deformable DETR [22] introduces a deformable attention mechanism that restricts attention computation to relevant spatial positions, greatly improving training efficiency and detection accuracy, especially for small and densely objects. Following this, DAB-DETR [23] and DINO [24] further refine query initialization and training dynamics, respectively, thus improving convergence speed and robustness.

Stemming from these foundational approaches, the Baidu team proposes Real-Time Detection Transformer(RT-DETR), a real-time end-to-end detection model that eliminates the inference latency caused by Non-Maximum Suppression. While maintaining high detection accuracy, RT-DETR achieves efficient real-time performance, demonstrating significant potential in practical applications—particularly in UAV-based object detection tasks, where rapid response is crucial. However, in aerial vision, small objects typically occupy very few pixels, and their discriminability heavily relies on contextual information. Relying solely on local features makes it challenging to fully model the semantic relationships between objects.Although recent works have explored dynamic feature modulation strategies[25] to enhance representation capability in general visual tasks, these mechanisms have not been fully adapted to the specific challenges of extreme scale variation in aerial imagery.Therefore, effectively enhancing shallow semantic perception and modeling complex inter-object relationships while maintaining a lightweight structure remains a critical challenge in current UAV-OD research.

Regarding these issues, we construct a Lightweight Receptive Enhancement Network for small objects that simultaneously expands the receptive field and preserves critical feature details, a Scale-Robust Feature Interaction module that dynamically retains multi-scale features, and an adaptive loss function that focuses on hard-to-match samples.The experimental results demonstrate the superiority of our method through the synergistic effect of these three aspects.

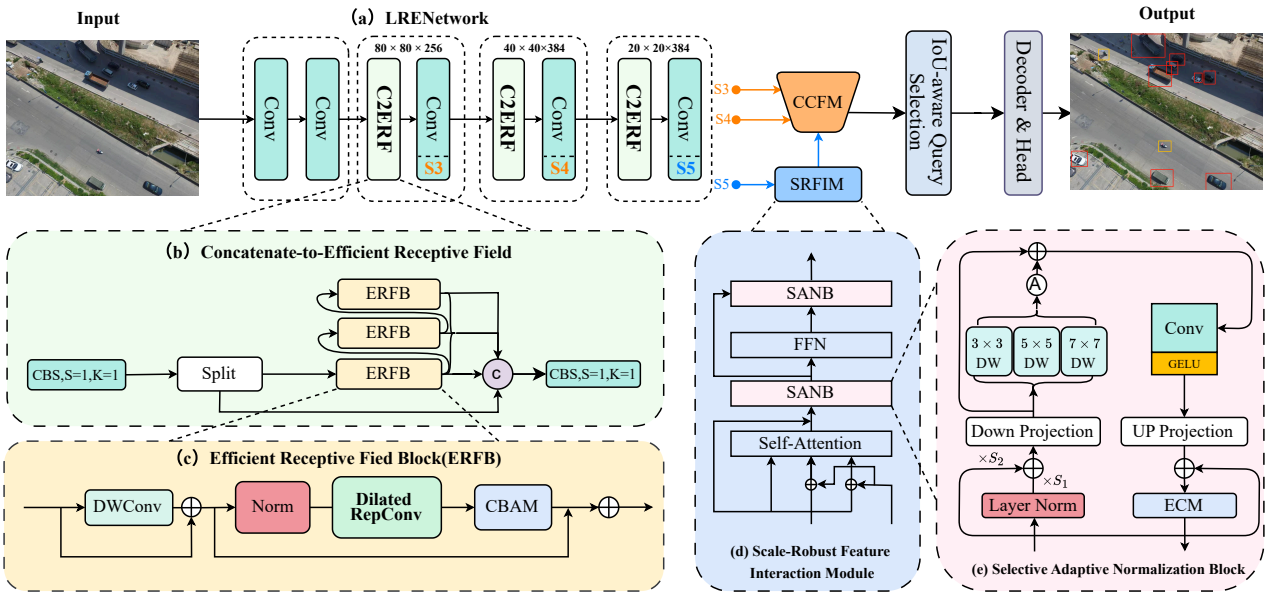


Figure 2. Overview of LME-DETR: (a) Lightweight Receptive Enhancement Network (LRENetwork), which includes the Concatenate-to-Efficient Receptive Field (C2ERF) and Efficient Receptive Field Block (ERFB), detailed in Section 3.1. (d) Scale-Robust Feature Interaction Module (SRFIM), and (e) Selective Adaptive Normalization Block (SANB), both detailed in Section 3.2. CCFM represents the Cross-Scale Feature Fusion Module.

## 3. Method

### 3.1. Lightweight Receptive Enhancement Network

For object detection, particularly for applications like Unmanned Aerial Vehicle (UAV) vision, standard backbones such as ResNet[26], exhibit serious limitations. The traditional hierarchical structure, built on successive convolution and pooling layers, aggressively downsamples feature maps to build semantic depth. This process, however, leads to a significant loss of spatial resolution and fine-grained details, which is particularly detrimental to detecting small objects that occupy only a few pixels. The network's ability to perceive these targets will weaken at deeper layers, thereby affecting the detection accuracy.

Additionally, small objects contain fewer intrinsic features, which requires the model to highly rely on surrounding context during recognition[27]. However, the receptive fields of Convolutional Neural Networks(CNNs)[28], expanded through stacked small kernels, are inherently dense and homogeneous. Although they are effective for locally clustered patterns, they cannot model the sparse, long-range dependencies critical for distinguishing a small object from its vast and often complex background[29]. However, simply increasing the kernel size is a flawed solution. It not only incurs prohibitive computational costs but also fails to alter the basic characteristics of the receptive fields, often blurring the very details it aims to contextualize[30]. In a network lacking the ability to build sparse, long-range connections, the pixels of a small object remain isolated and meaningless noise.

To this end, a paradigm shift is needed from passive information preservation to active contextual construction. The operation nature of standard convolution treats all spatial regions with equal importance resulting in the faint signals of small objects to be submerged by irrelevant background. An effective backbone must not only observe a wider region but also actively and selectively establish connections, focusing computational resources on the most remarkable information. Therefore, an advanced architecture is required to capture sparse long-range contextual dependencies while preserving high-resolution details within a lightweight framework suitable for real-time applications, and adaptively prioritize the discriminative features.

To overcome the challenges of information loss and inadequate contextual modeling in traditional CNNs for small object detection, We propose the Lightweight Receptive Enhancement Network named LRENetwork shown in Fig. 2(a) to achieve the paradigm shift from passive information aggregation to active contextual construction. LRENetwork fundamentally reshapes the feature extraction by alternately stacking standard downsampling convolutional layers with our pro-

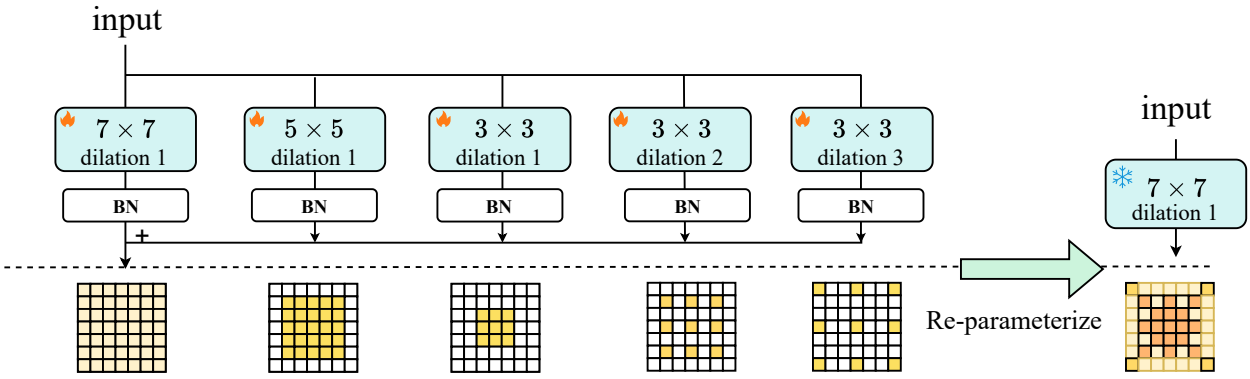


Figure 3. The Dilated RepConv is unfolded using an example with K=7. It employs multiple dilated convolutional branches to approximate a non-dilated large-kernel convolution, and additional dilated layers can be incorporated to support a larger effective kernel size K.

posed Context-aware Efficient Receptive Field Module (C2ERF). The convolutional layers are responsible for efficient spatial downsampling, and C2ERF actively perform feature refinement and contextual enhancement at each scale. This hierarchical processing generates a powerful feature pyramid, and we ultimately select the outputs from its three key stages (S3,S4,S5) to serve as multi-scale inputs for the subsequent hybrid encoder, providing a robust foundation for rich object perception and structural semantic modeling.

The heart of our LRENetwork is the designed C2ERF module (Fig.2(b)), whose internal structure is critical to achieving the dual goals of active construction and detail preservation. C2ERF employs an efficient dual-path parallel architecture. After an initial convolution, input features are bifurcated into two paths. A residual path acts as a shortcut to directly preserve the original features, ensuring that the fine-grained spatial details crucial for small objects are not diluted during deep propagation. Concurrently, a primary path feeds the features into a series of cascaded Efficient Receptive Field Blocks named ERFB shown in Fig.2(c), which is the key of our active contextual construction strategy. The outputs from these two paths are then fused through a residual connection and integrated by a final convolutional layer. This design enables C2ERF to significantly enhance feature representation while maintaining minimal computational overhead, achieving a balance between performance and efficiency.

Efficient Receptive Field Block is designed to directly address the fundamental limitation of conventional convolutions—their uniform and non-selective receptive fields. It consists of three key components: a 3x3 depthwise convolution (DWConv) for efficient local spatial modeling; a Dilated RepConv(DRC)[31]for building sparse, long-range contextual connections; and a Convolutional Block Attention Module(CABM)[32] for adaptively selecting and enhancing salient information.

Small objects occupy only a limited number of pixels in an image. Motion blur, occlusions, or limited resolution may result in incomplete appearance and weak discriminative features. As a result, their accurate recognition in UAV often relies heavily on rich contextual information. The semantic cues required for correct classification or localization are not necessarily concentrated around the object itself, but are instead sparsely distributed across a broader spatial region. Therefore, it is essential to design a model capable of capturing such sparse long-range dependencies while maintaining computational efficiency.

Inspired by the convolutional enhancement strategy proposed by Xiaohan Ding et al. [33], which advocates combining parallel small-kernel convolutions with large-kernel convolutions to enhance fine-grained pattern modeling, we put in the Dilated RepConv(DRC). DRC adopts a training-inference decoupling mechanism and employs multiple small-kernel branches with varying dilation rates during training to capture multi-scale and sparse receptive field patterns.(see Fig. 3 e.g., $k = \{5, 3, 3, 3\}$ and $r = \{1, 1, 2, 3\}$)

Specifically, the branch with $r = 1$ focuses on local fine-grained structure modeling, while those with $r > 1$ contribute to capturing sparse and spatially distant semantic responses, which are critical for small object detection. Each branch independently performs convolution and normalization, and their outputs are aggre-

gated in a weighted fashion to form a unified, context-enhanced feature representation. This promotes improved semantic awareness and multi-scale sensitivity.

After training, all dilated branches are structurally reparameterized[34] into a single dense convolutional layer for efficient inference. Each dilated kernel is first transformed into an equivalent non-dilated sparse kernel with effective size $(k-1)r+1$ , followed by batch normalization fusion. The resulting kernels are zero-padded to match target dimensions and summed to obtain a unified large-kernel convolution.

However, although DRC effectively models sparse contextual patterns, it still applies a uniform convolutional operation across all spatial regions and channels. This lacks explicit modeling of region-specific importance, potentially causing the crucial semantic cues of small targets to be overwhelmed by irrelevant background activations. To address this limitation, we introduce the lightweight CBAM attention mechanism immediately after the DRC. By sequentially applying channel and spatial attention, CBAM adaptively recalibrates the feature map. It highlights informative channels strongly correlated with target semantics and emphasizes spatial regions likely to contain discriminative object patterns. This dual-attention mechanism guides the model to focus its computational resources on the most critical information while suppressing redundant or distracting background noise, thereby realizing the active, efficient, and selective feature enhancement.

### 3.2. Scale-Robust Feature Interaction Module

The Attention-based Intra-scale Feature Interaction (AIFI) module in RT-DETR encoder is to enhance the representation of high-level semantic features. By performing self-attention independently on the high-level S5 feature layer, AIFI effectively captures semantic correlations among objects. This approach avoids redundant processing of low-level features, thereby reducing computational overheads and improving detection efficiency.

However, AIFI is constructed upon a standard Transformer encoder architecture, which consists of multi-head self-attention and a feed-forward network, utilizing a conventional Add & Norm layer for feature fusion and normalization. This static normalization strategy demonstrates significant limitations in multi-scale detection scenarios, a challenge particularly prevalent in Unmanned Aerial Vehicle (UAV) aerial imagery. The indiscriminate processing of features by the traditional Add & Norm layer may suppress the weak semantic representations of small targets, resulting in information being overshadowed by large, dominant objects and ultimately degrading small object detection performance.

To address the issue, we construct the Scale-Robust Feature Interaction Module (SRFIM) shown in Fig. 2(d), a novel architecture designed to enhance multi-scale modeling capabilities and improve the identification and preservation of small target features. The innovative Selective Adaptive Normalization Block (SANB) shown in Fig. 2(e) and incorporated within SRFIM, integrates three key mechanisms: (1) a weak feature protection mechanism that employs dynamic normalization parameter estimation to prevent dominant features from overwhelming small-target semantics. (2) a multi-scale structural modeling strategy that establishes diverse feature pathways to improve adaptability to objects of varying sizes and (3) a channel-wise selection mechanism that enhances the semantic expressiveness of the fused multi-scale features.

Grounded in the principle of structural decoupling, SANB separates the normalization from the residual information fusion pathway and employs learnable scaling factors to dynamically weight these two components, thereby improving the preservation of subtle semantic information. Given the input feature $x_0 \in \mathbb{R}^{H\times W\times C_{in}}$, it is first dynamically adjusted by a Layer Normalization (LN) followed by two learnable scaling weights $S_1$ and $S_2$:

$$x_{\text{norm}} = \text{LayerNorm}(x_0)\cdot S_1 + x_0 \cdot S_2, \tag{1}$$

where, $S_1 \in \mathbb{R}^{C_{in}\times 1\times 1}$ is a learnable scaling parameter initialized to $1\times 10^{-6}$, while $S_2 \in \mathbb{R}^{C_{in}\times 1\times 1}$ is a learnable weight initialized to 1. This mechanism effectively balances normalization with the retention of original semantics, aiding in the preservation of weak but crucial features during the fusion.

To enhance the network's adaptability to varying object sizes, the normalized feature $x_{\text{norm}}$ is first compressed along the channel dimension to a fixed size of $C_{\text{dim}} = 64$ via a linear projection, denoted as $x_{\text{down}} = DP(x_{\text{norm}})$. This compressed representation is then simultaneously passed through three parallel depthwise separable convolutional branches, each with distinct kernel sizes:

$$f_{dw}^{(i)} = DWConv_{K_i}(x_{\text{down}}), \quad i = 1,2,3, \tag{2}$$

where $K_i \in \{3,5,7\}$ represents the receptive field scale of each branch. These multi-kernel branches allow the model to capture diverse contextual patterns across varying spatial ranges.

The outputs from the three branches are subsequently averaged and aligned in channel space through a point-wise convolution, followed by a residual skip connection to preserve the original low-level features.

The result is further activated by a GeLU function and upsampled via a learned projection layer:

$$x = x_0 + U^l \cdot \sigma(f_{pw}(f_{dw}(D^l(x_{\text{norm}})))), \tag{3}$$

where $D^l$ and $U^l$ denote the down- and up-projection layers in the $l^{\text{th}}$ SANB, and $\sigma$ is the GeLU nonlinearity. This hierarchical design ensures rich multi-scale feature modeling while maintaining computational efficiency.

To further enhance the semantic expressiveness of the fused multi-scale features, we append a lightweight channel attention mechanism after the up-projection operation. Specifically, we adopt the Efficient Channel Attention module (ECA)[35], a parameter-efficient yet effective attention design that adaptively emphasizes informative channels while suppressing less relevant ones.

Unlike conventional channel attention mechanisms such as Squeeze-and-Excitation (SE) [36] that rely on full channel compression followed by multi-layer perceptrons (MLPs), ECA avoids dimensionality reduction and instead captures local cross-channel interactions through a 1D convolution with a carefully selected kernel size. This design eliminates the need for explicit channel squeezing, allowing ECA to better preserve channel-wise information continuity and avoid information bottlenecks.

Formally, given an input feature map $F \in \mathbb{R}^{C\times H\times W}$, the ECA module first generates channel weights by applying Global Average Pooling (GAP), followed by a 1D convolution ($\text{Conv1D}_k$) and a sigmoid activation ($\sigma$). These weights are then applied back to the input feature map via channel-wise multiplication. This entire sequence can be compactly expressed as:

$$F' = \sigma(\text{Conv1D}_k(\text{GAP}(F))) \odot F, \tag{4}$$

where $\odot$ denotes the channel-wise product, which broadcasts the generated channel weight vector across the spatial dimensions of $F$.

We apply ECA after the up-projection operation rather than before, ensuring that the attention operates in a semantically richer feature space with restored dimensionality. This ordering is critical, as it allows the attention mechanism to exploit more discriminative cues for highlighting target objects, especially when the feature stems from the multi-scale receptive field encoding established by our C2ERF module. ECA introduces negligible computational overhead and avoids dimensional mismatch when used in residual branches, making it highly suitable for our lightweight design.

### 3.3. Exponential Moving Average Slide Varifocal Loss

The training process for dense small object detection is hindered by two primary issues related to the quality of positive samples, which is typically measured by the Intersection-over-Union (IoU). The IoU quantifies the overlap between a predicted bounding box $B_p$ and a ground-truth box $B_{gt}$, and is defined as:

$$\text{IoU}(B_p, B_{gt}) = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})}. \tag{5}$$

First, although positive samples are numerous, their IoU scores are highly imbalanced. Second, small objects, due to their low matching quality (i.e., low IoU scores), are often misinterpreted as noisy samples, which degrades model performance. While Varifocal Loss (VFL) [37] addresses class imbalance by weighting samples based on their IoU, its static modulation structure is a significant limitation. It applies a fixed weighting strategy regardless of the training stage, making it insufficiently adaptive, especially in the early phases when IoU distributions are unstable and small objects are poorly matched.

To overcome this rigidity, we propose the Exponential Moving Average Slide Varifocal Loss (EMA-SVFL), a novel loss function that introduces a dynamic, adaptive training objective. The core principle is twofold: first, we establish a dynamic baseline of matching quality by tracking a running average of positive sample IoUs. Second, we leverage this evolving baseline to implement a piecewise weighting strategy that intensifies supervision on hard-to-match samples (those with IoU below the baseline) and reduces focus on already well-matched samples.

Specifically, at each training step $t$, we first estimate the global IoU center, denoted as $\mu_t$. This is achieved by updating the previous center $\mu_{t-1}$ (initialized as $\mu_0 = 0.5$) with the average IoU of positive samples in the current batch, $q_t$. The update rule is formulated as:

$$\mu_t = d_t \cdot \mu_{t-1} + (1 - d_t) \cdot q_t, \tag{6}$$

where the dynamic momentum term $d_t = \lambda(1 - e^{-t/\tau})$ controls the update rate. In early training ($t \to 0$), $d_t$ is small, making $\mu_t$ highly responsive to the current batch's statistics. As training progresses ($t \to \infty$), $d_t$ approaches $\lambda$, ensuring a stable and smooth estimation.

Based on this dynamic center $\mu_t$, we introduce a three-phase modulation function, $\omega(q, \mu_t)$, to adaptively re-weight each foreground sample according to its relative matching quality:

$$\omega(q, \mu_t) = \begin{cases} 1 + d_t, & \text{if } q \le \mu_t - \delta \\ 1, & \text{if } \mu_t - \delta < q < \mu_t + \delta \\ 1 - d_t, & \text{if } q \ge \mu_t + \delta \end{cases} \tag{7}$$

Here, $\delta$ is a hyperparameter defining the boundary of the "medium-quality" region (we set $\delta = 0.1$). This function dynamically partitions positive samples into 'hard', 'medium', and 'easy' categories relative to the current training state, thus focusing the model's learning capacity where it is most needed.

We integrate this adaptive weighting mechanism into the original VFL formulation. The complete EMA-SVFL is defined as the sum of the foreground loss $\mathcal{L}_{\text{pos}}$ and the background loss $\mathcal{L}_{\text{neg}}$:

$$\mathcal{L}_{\text{EMA-SVFL}} = \frac{1}{N_{\text{pos}}} \sum_{i \in \mathcal{P}} \mathcal{L}_{\text{pos}}(p_i, q_i) + \frac{1}{N_{\text{neg}}} \sum_{j \in \mathcal{N}} \mathcal{L}_{\text{neg}}(p_j), \tag{8}$$

where $\mathcal{P}$ and $\mathcal{N}$ are the sets of positive and negative samples, with sizes $N_{\text{pos}}$ and $N_{\text{neg}}$, respectively. The individual loss terms are defined as:

$$\mathcal{L}_{\text{pos}}(p, q) = -\omega(q, \mu_t) \cdot q^{\gamma} \log(p), \tag{9}$$

$$\mathcal{L}_{\text{neg}}(p) = -\alpha p^{\gamma} \log(1 - p). \tag{10}$$

In these equations, $p$ is the predicted classification confidence, $q$ is the IoU between the predicted and ground-truth boxes, and $\alpha$ and $\gamma$ are the standard VFL hyperparameters. This design offers several key advantages: it adaptively enhances supervision for low-IoU samples without requiring manually-tuned IoU thresholds for loss modulation; it introduces no learnable parameters, making it a lightweight and universally pluggable module; and it demonstrates greater stability during the volatile early stages of training, facilitating faster convergence and superior final performance.

## 4. Experiments

### 4.1. Datasets and Experimental Configuration

To verify the effectiveness of our proposed LME-DETR, we conduct experiments on two representative UAV vision datasets: VisDrone2019 [16] and UAVVaste[17]. VisDrone2019 contains 6,471 training images, 548 validation images, and 3,190 test images. All images were captured by UAVs at various geographic locations and flying altitudes. Each image is annotated with bounding boxes covering ten predefined object categories, including pedestrian, person, car, van, bus, truck, motorbike, bicycle, awning-tricycle, and tricycle. UAVVaste is a dataset specifically designed for aerial litter detection. It consists of 772 images and 3,716 manually annotated instances of waste objects distributed across both urban and natural environments, such as streets, parks, and grasslands.

All models are trained on an NVIDIA RTX 5060 Ti GPU. For a total of 300 epochs with a batch size of 4 respectively. An early stopping mechanism is adopted to prevent overfitting, with the patience value set to 20. We use AdamW [38] as the optimizer , with a learning rate of $1 \times 10^{-4}$ and a momentum of 0.9.

The Average Precision (AP) is used for evaluating the model's detection performance on a single category, defined as the area under the Precision-Recall curve, reflecting the model's average performance across different confidence thresholds. The calculation formula is as follows:

$$AP = \int_0^1 P(R) dR, \tag{11}$$

where $P$ denotes Precision and $R$ denotes Recall. In this study, we adopt $\text{AP}_{50}$ as the primary evaluation metric, which represents the Average Precision calculated at an IoU(Formula (5)) threshold of 0.5, used to assess the model's fundamental object detection capability.

To further analyze the detection performance of objects of different scales, we record $\text{AP}_S$, $\text{AP}_M$, and $\text{AP}_L$, which correspond to small, medium and large objects, respectively. These metrics are defined based on the object area within the image: small objects refer to those with areas smaller than $32^2$ pixels, medium objects between $32^2$ and $96^2$, and large objects greater than $96^2$. In UAV-based detection scenarios, small objects are especially prevalent and often suffer from appearance degradation due to resolution constraints, motion blur, and occlusion, making $\text{AP}_S$ a critical metric for performance assessment. Altogether, these metrics provide a multi-dimensional evaluation of both localization accuracy and detection robustness across various object sizes and scene complexities

### 4.2. Experimental results and analysis

#### 4.2.1 Results on Visdrone Dataset

We first conduct a comprehensive evaluation on the challenging VisDrone2019 benchmark, which is characterized by densely distributed, small-scale, and highly cluttered aerial targets. The results in Table 1 demonstrate that LME-DETR achieves an outstanding balance between detection accuracy, computational efficiency, and scale robustness.

Compared with the direct baseline RT-DETR-R18, our method delivers a substantial performance improvement. The overall AP increases by 11.6 percent, and $\text{AP}_{50}$ rises by 9.0 percent, reflecting a clear enhancement in both general precision and high-confidence detection capability. At the same time, LME-DETR reduces the parameter count by about thirty percent, the computational cost by nineteen percent, and achieves a twenty percent faster inference speed. This establishes our model as a more compact

Table 1. Comparison of representative object detectors with comparable parameter scales on VisDrone2019. * denotes the best performance and the underscore '_' denotes the second-best performance.

| Model | Publication | Input Shape | GFlops | Params | FPS | AP | $AP_{50}$ |
|---|---|---|---|---|---|---|---|
| Two-stage Methods | | | | | | | |
| Faster-RCNN[18] | TPAMI 2017 | (768, 1344) | 208G | 41.39M | 24 | 19.4 | 32.9 |
| Cascade-RCNN[39] | CVPR 2018 | (768, 1344) | 236G | 69.29M | 18 | 19.7 | 32.6 |
| One-stage Methods | | | | | | | |
| YOLOv8s[40] | ADICS 2024 | (640, 640) | 28G | 11.13M | 131 | 27.3 | 40.7 |
| YOLOv8m[40] | ADICS 2024 | (640, 640) | 78G | 25.85M | 83 | 27.1 | 43.2 |
| YOLOv10s[41] | NeurIPS 2024 | (640, 640) | 21G | 7.22M | 135 | 17.9 | 42.3 |
| YOLOv10m[41] | NeurIPS 2024 | (640, 640) | 58G | 15.32M | 77 | 29.5 | 44.5 |
| YOLOv12s[42] | arXiv 2025 | (640, 640) | 21G | 9.23M | 132 | 27.6 | 41.2 |
| YOLOv12m[42] | arXiv 2025 | (640, 640) | 67G | 20.11M | 67.5 | 29.2 | 43.6 |
| FBRT-YOLO[43] | AAAI 2025 | (640, 640) | 119G | 14.60M | 70 | 29.7 | <u>47.7</u> |
| End-to-End Methods | | | | | | | |
| DETR[21] | ECCV 2020 | (640, 640) | 60G | 187M | 38 | 24.1 | 40.1 |
| Deformable DETR[22] | ICLR 2020 | (640, 640) | 173G | 40M | 62 | 27.1 | 42.2 |
| RT-DETR-R18[15] | CVPR 2024 | (640, 640) | 57G | 20M | <u>183</u> | 28.5 | 44.6 |
| D-Fine-M[44] | ICLR 2025 | (640, 640) | 57G | 19.9M | 165 | 33.9* | 41.6 |
| LME-DETR (Ours) | - | (640, 640) | 46G | 14M | 220* | <u>31.8</u> | 48.6* |

Table 2. COCO Size-based AP Metrics ($AP_S, AP_M, AP_L$) on the VisDrone and UAVVaste Datasets. * denotes the best performance and the underscore '_' denotes the second-best performance.

| | VisDrone | | | UAVVaste | | |
|---|---|---|---|---|---|---|
| Model | $AP_S$ | $AP_M$ | $AP_L$ | $AP_S$ | $AP_M$ | $AP_L$ |
| Two-stage Methods | | | | | | |
| Faster-RCNN | 9.5 | 30.6 | 42.9 | 32.1 | 58.3 | 69.7 |
| Cascade-RCNN | 9.9 | 30.7 | 40.6 | 29.8 | 55.2 | 67.4 |
| One-stage Methods | | | | | | |
| YOLOv8s | 8.8 | 28.1 | 40.2 | 33.5 | 60.1 | 70.5 |
| YOLOv8m | 9.0 | 29.4 | 41.7 | 34.2 | 61.5 | 72.3 |
| YOLOv10s | 9.2 | 28.9 | 40.1 | 34.1 | 60.8 | 70.2 |
| YOLOv10m | 9.7 | 30.0 | 41.4 | 34.8 | 62.1 | 71.8 |
| YOLOv12s | 9.1 | 28.5 | 37.5 | 34.7 | 61.9 | 69.0 |
| YOLOv12m | 9.4 | 29.8 | 38.6 | 35.4 | 63.2 | 70.1 |
| FBRT-YOLO | 9.4 | 30.9* | 42.1 | 35.1 | 64.7 | 73.2 |
| End-to-End Methods | | | | | | |
| DETR | 11.1 | 23.5 | 40.6 | 31.8 | 57.9 | 68.4 |
| Deformable DETR | 12.4 | 25.7 | 42.4 | 34.5 | 60.3 | 71.6 |
| RT-DETR-R18 | 11.3 | 27.5 | 42.3 | 35.8 | 64.8 | 75.7 |
| D-Fine-M | <u>12.8</u> | 29.5 | <u>43.1</u> | <u>36.5</u> | <u>66.2</u> | <u>77.0</u> |
| LME-DETR (Ours) | 13.2* | <u>30.7</u> | 43.7* | 37.4* | 67.8* | 78.9* |

yet stronger alternative within the real-time DETR family.

Beyond the baseline comparison, LME-DETR also exhibits superiority against other state-of-the-art methods. In contrast to the recently proposed D-Fine-M, which attains a comparable average precision, our method achieves a much higher $AP_{50}$, surpassing it by 16.8 percent, which indicates stronger detection reliability and more stable object localization under identical resource constraints. When compared with the representative YOLOv12m, our model achieves an 8.9 percent increase in AP, while maintaining a model size smaller by nearly one third. This demonstrates that the proposed design not only enhances accuracy

but also provides a favorable accuracy-to-cost trade-off, outperforming both transformer-based and convolutional architectures.

To further analyze the source of these improvements, we investigate the performance across different object scales, as summarized in Table 2. The VisDrone2019 dataset contains a large proportion of small targets, which represents a major bottleneck for most detectors. Our method demonstrates clear dominance in this regime. LME-DETR achieves the highest $AP_S$ of 13.2, significantly outperforming RT-DETR-R18 by 16.8 percent, the specialized small-object detector FBRT-YOLO by 40.4 percent, and even D-Fine-M by 3.1 percent. These improvements highlight the strong capability of LME-DETR to preserve fine-grained visual cues and enhance small-object perception under complex aerial conditions.

In addition to the small-object improvements, LME-DETR maintains competitive results on medium and large targets, achieving the highest $AP_L$ of 43.7 among all compared methods. This consistent advantage across scales confirms that the proposed Lightweight Receptive Enhancement design effectively expands the receptive field without introducing excessive computational overhead. Consequently, LME-DETR not only excels at recognizing small, densely packed targets but also maintains stable performance on large-scale instances, revealing its robustness and scalability in real-world UAV perception scenarios.

#### 4.2.2 Results on UAVVaste Dataset

To further validate the generalization and robustness of our method, we conduct experiments on the UAVVaste dataset, which features large scene diversity, complex illumination conditions, and frequent occlusions. The results summarized in Table 3 demonstrate that LME-DETR maintains its superior balance between accuracy and efficiency when transferred to a new UAV scenario without additional fine-tuning.

Among all compared approaches, LME-DETR achieves the highest overall AP of 37.6 and the top $AP_{50}$ of 80.5, indicating remarkable detection reliability and stable localization quality. In contrast, the transformer-based models such as D-Fine-M and RT-DETR-R18 rely on deeper backbones and heavier attention computation, while the one-stage YOLOv12m still struggles with small and mid-scale target recall. Our model bridges these gaps by combining the interpretability of end-to-end Transformers with the compactness of CNN-based architectures, leading to a more robust and deployable UAV detector.

The multi-scale breakdown reported in Table 2 further confirms the effectiveness of our design. LME-DETR consistently achieves the highest scores for small, medium, and large objects, demonstrating balanced representation capability across different spatial resolutions. Notably, its advantage on small and medium targets indicates that the Lightweight Multi-scale Enhancement mechanism successfully strengthens feature aggregation and spatial awareness under real-world aerial conditions.

Beyond the numerical gains, the consistent performance trends observed across VisDrone2019 and UAVVaste reflect the cross-dataset stability of our architecture. The model not only adapts to varying flight altitudes and viewpoints but also preserves robust performance in scenes with heavy clutter or illumination imbalance. This cross-domain consistency confirms that LME-DETR generalizes effectively to unseen UAV environments, providing a solid foundation for real-time aerial perception applications.

### 4.3. Ablation Study

To understand the individual and combined contributions of each proposed component, we perform an incremental ablation study on the VisDrone2019 dataset, as presented in Table 4. Starting from the RT-DETR baseline, we sequentially integrate the LRENetwork, SRFIM, and EMA-SVFL modules to examine how each design improves detection performance across different object scales.

The first modification replaces the baseline backbone with our LRENetwork, resulting in a clear performance gain, particularly for small and large targets. The value of $AP_S$ increases from 11.3 to 12.1, while $AP_L$ improves from 41.3 to 42.7. This improvement stems from the LRENetwork's ability to expand the model's effective receptive field through reparameterizable dilated convolutions. The resulting enhancement demonstrates that a well-calibrated receptive field is fundamental to balancing fine-grained detail extraction and global scene understanding in aerial imagery.

Integrating the SRFIM module further amplifies the network's discriminative power, pushing the small-object accuracy to an $AP_S$ of 12.8. This gain highlights the importance of cross-scale semantic reinforcement, where SRFIM bridges the gap between deep semantic abstractions and shallow spatial features. By promoting bidirectional information flow across feature hierarchies, SRFIM helps the model maintain spatial precision while gaining contextual awareness, which is crucial for detecting densely packed and visually ambiguous UAV targets.

Finally, the inclusion of the EMA-SVFL loss function drives the model to its best performance, achieving

Table 3. Comparison of representative object detectors with comparable parameter scales on UAVVaste. * denotes the best performance and the underscore '_' denotes the second-best performance.

| Model | Publication | Input Shape | GFlops | Params | FPS | AP | $AP_{50}$ |
|---|---|---|---|---|---|---|---|
| Two-stage Methods | | | | | | | |
| Faster-RCNN | TPAMI 2017 | (768, 1344) | 208G | 41.39M | 24 | 30.3 | 67.0 |
| Cascade-RCNN | CVPR 2018 | (768, 1344) | 236G | 69.29M | 18 | 27.8 | 65.0 |
| One-stage Methods | | | | | | | |
| YOLOv8s | ADICS 2024 | (640, 640) | 28G | 11.13M | 131 | 34.5 | 69.8 |
| YOLOv8m | ADICS 2024 | (640, 640) | 78G | 25.85M | 83 | 35.7 | 71.7 |
| YOLOv10s | NeurIPS 2024 | (640, 640) | 21G | 7.22M | 135 | 34.9 | 68.5 |
| YOLOv10m | NeurIPS 2024 | (640, 640) | 58G | 15.32M | 77 | 35.8 | 70.6 |
| YOLOv12s | arXiv 2025 | (640, 640) | 21G | 9.23M | 132 | 36.1 | 71.2 |
| YOLOv12m | arXiv 2025 | (640, 640) | 67G | 20.11M | 67.5 | 37.4 | 73.4 |
| FBRT-YOLO | AAAI 2025 | (640, 640) | 119G | 14.60M | 70 | 36.9 | <u>78.6</u> |
| End-to-End Methods | | | | | | | |
| DETR | ECCV 2020 | (640, 640) | 60G | 187M | 38 | 34.8 | 71.6 |
| Deformable DETR | ICLR 2020 | (640, 640) | 173G | 40M | 62 | 37.1 | 74.7 |
| RT-DETR-R18 | CVPR 2024 | (640, 640) | 57G | 20M | <u>183</u> | 36.3 | 72.6 |
| D-Fine-M | ICLR 2025 | (640, 640) | 57G | 19.9M | 165 | <u>37.5</u> | 74.2 |
| LME-DETR (Ours) | - | (640, 640) | 46G | 14M | 220* | 37.6* | 80.5* |

Table 4. Ablation study of each component in LME-DETR on the VisDrone dataset. The experiment starts with the RT-DETR baseline and progressively adds our proposed modules.

| Baseline | LRENetwork | SRFIM | EMA-SVFL | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|
| ✓ | | | | 11.3 | 30.5 | 41.3 |
| ✓ | ✓ | | | 12.1 | 29.9 | 42.7 |
| ✓ | ✓ | ✓ | | 12.8 | 30.6 | 43.2 |
| ✓ | ✓ | ✓ | ✓ | 13.2 | 30.7 | 43.7 |

the highest values across all scale metrics. This improvement validates the role of adaptive supervision in enhancing robustness and generalization. By dynamically emphasizing hard or underrepresented samples—such as small, occluded, or low-contrast objects—and stabilizing gradient updates through an exponential moving average, EMA-SVFL effectively reallocates the learning focus toward the most informative examples. The result is a detector that not only achieves higher accuracy but also exhibits stronger consistency under challenging aerial conditions.

### 4.4. Visualization Experiments

Beyond quantitative gains, we further visualize how each design in LME-DETR improves perception behavior across diverse UAV scenes. To this end, Fig. 4 presents a unified qualitative comparison on both the VisDrone and UAVVaste datasets, providing a comprehensive perspective on spatial attention and detection robustness.

On the left side of Fig. 4, we analyze the activation heatmaps derived from VisDrone. These results expose a clear perceptual disparity between the baseline RT-DETR and our proposed LME-DETR. The baseline suffers from fragmented and diffuse attention, leaving large unresponsive regions over dense or overlapping targets, and occasionally activating irrelevant background textures such as roads or vegetation. In contrast, LME-DETR yields compact and high-intensity activation clusters accurately aligned with true targets, while maintaining strong suppression over background noise. This improvement originates from two key architectural innovations. The Lightweight Receptive Enhancement Network (LRENet) expands the effective receptive field to capture richer contextual cues without extra computational cost, and the SRFIM module facilitates scale-aware feature validation by allowing global semantics to guide local structural refinement. Together, these mechanisms enable the model to perceive fine-grained details while maintaining coherent global understanding.

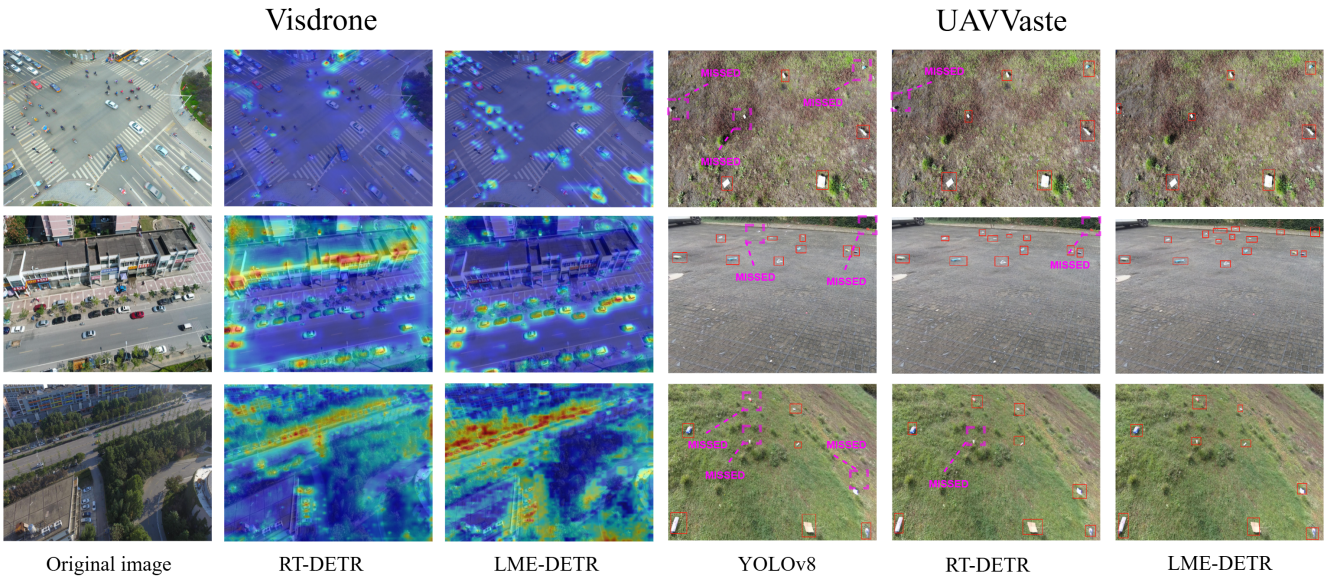


Figure 4. Comprehensive visualization on the VisDrone (left) and UAVVaste (right) datasets. The VisDrone heatmaps compare the spatial attention distributions of RT-DETR and our proposed LME-DETR. The UAVVaste results further highlight detection robustness in real-world UAV imagery: baseline models such as YOLOv8 and RT-DETR exhibit frequent missed detections (marked as "missed") and false activations.

## 5. Conclusion

In this paper, we propose LME-DETR, an efficient end-to-end object detector specifically engineered to address the critical challenges of UAV-based visual perception.The Lightweight Receptive field Enhancement Network (LRENetwork) is designed to provide robust contextual modeling while maintaining a compact footprint. The Scale-Robust Feature Interaction Module (SRFIM) is put in to foster effective cross-scale feature fusion, significantly enhancing the model's robustness to objects of varying sizes. Our proposed EMA-SlideVarifocalLoss (EMA-SVFL) intensifies supervision on difficult-to-detect small objects by adaptively shifting the optimization focus. Extensive experiments on UAV benchmarks confirm that our LME-DETR achieves a state-of-the-art balance between accuracy and efficiency. These contributions establish LME-DETR not only as a novel framework but as a practical and powerful solution ready for deployment in real-world UAV applications.

## Acknowledgement

This work was supported by the Shanghai Municipal Education Commission Artificial Intelligence Plan (Z2024-119), the Innovation Special Fund Project in Shanghai University of Electric Power (Grant No. X202511010), and the Industry–University Collaborative Research Project "Large Language Model–Based Intelligent Operation and Maintenance Knowledge Platform for Power Grid Equipment" (No. H2025-220).

## References

[1] B. Wang, W. Li, B. Zhang, and Y. Liu. Joint response and background learning for uav visual tracking. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 455–462, 2024. 1

[2] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2874–2883, 2016. 2

[3] Z. Du, J. Yin, and J. Yang. Expanding receptive field yolo for small object detection. Journal of Physics: Conference Series, 1314(1):012202, oct 2019. 2

[4] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell. Understanding convolution for semantic segmentation. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1451–1460, 2018. 2

[5] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. ArXiv, abs/1706.05587, 2017. 2

[6] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, Computer Vision – ECCV 2018, pages 833–851, Cham, 2018. Springer International Publishing. 2

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–9, 2015. 2

[8] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9):1904–1916, 2015. 2

[9] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5987–5995, 2017. 2

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 936–944, 2017. 2

[11] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8759–8768, 2018. 2

[12] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10778–10787, 2020. 2

[13] G. Ghiasi, T.-Y. Lin, and Q. V. Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7029–7038, 2019. 2

[14] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 2

[15] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen. Detrs beat yolos on real-time object detection. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 16965–16974, 2024. 2, 9

[16] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling. Detection and tracking meet drones challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(11):7380–7399, 2021. 2, 8

[17] M. Kraft, M. Piechocki, B. Ptak, and K. Walas. Autonomous, onboard vision-based trash and litter detection in low altitude aerial images collected by an unmanned aerial vehicle. Remote Sensing, 13(5), 2021. 2, 8

[18] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6):1137–1149, 2017. 3, 9

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2980–2988, 2017. 3

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2999–3007, 2017. 3

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, Computer Vision – ECCV 2020, pages 213–229, Cham, 2020. Springer International Publishing. 3, 9

[22] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159, 2020. 3, 9

[23] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In International Conference on Learning Representations, 2022. 3

[24] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605, 2022. 3

[25] D. Yin, L. Hu, B. Li, Y. Zhang, and X. Yang. 5%> 100%: Breaking performance shackles of full fine-tuning on visual recognition tasks. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 20071–20081, 2025. 3

[26] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016. 4

[27] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2874–2883, 2016. 4

[28] D. J. Santry. Convolutional Neural Networks, pages 111–131. 2024. 4

[29] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7794–7803, 2018. 4

[30] X. Ding, X. Zhang, J. Han, and G. Ding. Scaling up your kernels to 31×31: Revisiting large kernel design in cnns. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11953–11965, 2022. 4

[31] X. Ding, Y. Zhang, Y. Ge, S. Zhao, L. Song, X. Yue, and Y. Shan. Unireplknet: A universal perception large-kernel convnet for audio, video, point cloud, time-series and image recognition. In 2024 IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), pages 5513–5524, 2024. 5

[32] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII, page 3–19, Berlin, Heidelberg, 2018. Springer-Verlag. 5

[33] X. Ding, X. Zhang, J. Han, and G. Ding. Scaling up your kernels to 31×31: Revisiting large kernel design in cnns. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11953–11965, 2022. 5

[34] X. Ding, Y. Guo, G. Ding, and J. Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 1911–1920, 2019. 6

[35] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 7

[36] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7132–7141, 2018. 7

[37] H. Zhang, Y. Wang, F. Dayoub, and N. Sünderhauf. Varifocalnet: An iou-aware dense object detector. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8510–8519, 2021. 7

[38] I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. 11 2017. 8

[39] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6154–6162, 2018. 9

[40] R. Varghese and S. M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), pages 1–6, 2024. 9

[41] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding. Yolov10: Real-time end-to-end object detection. ArXiv, abs/2405.14458, 2024. 9

[42] Y. Tian, Q. Ye, and D. Doermann. Yolov12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524, 2025. 9

[43] Y. Xiao, T. Xu, Y. Xin, and J. Li. Fbrt-yolo: Faster and better for real-time aerial image detection. Proceedings of the AAAI Conference on Artificial Intelligence, 39(8):8673–8681, Apr. 2025. 9

[44] Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, and F. Wu. D-fine: Redefine regression task in detrs as fine-grained distribution refinement, 2024. 9